Learning Where to Look: The Acquisition of Location Knowledge in Display-Based Interaction

ABSTRACT

The locations of interface objects (e.g., buttons, menu items) are central and necessary components to direct manipulation; to be used, these objects must be located, pointed at, and clicked on. Knowledge of the locations of sought-after objects can significantly reduce the visual search space and thus reduce performance times. Research indicates that people do indeed learn the locations of interface objects and use this location knowledge to improve performance. The question of how the impetus and opportunity for location learning change as a function of the cost structure of an interface was explored via a two-phased approach; the first part empirical, and the second analytical. The empirical phase was comprised of two experiments. The first experiment employed an incidental learning paradigm in which participants perform a search and select task and were subsequently forced to rely on their location knowledge. Experiment II used the same search and select task as Experiment I, but involved the collection of eye gaze data as a longitudinal and direct behavioral measure of location learning.

The results of these experiments were used to constrain the behavior of a computational cognitive model in the second phase of the project. The model, built using ACT-R/PM (Anderson, 1993; Anderson & Lebiere, 1998; Byrne & Anderson, 1998), interacts with the same experimental task as the participants. Guided by the assumption that participants acted rationally, seeking maximum gain at minimum cost, the model

provided a compelling and detailed account of key attributes of participant's behavior, from fine-grained components of interaction such as eye and mouse movements to higher order measures such as performance time. An analysis of the underlying assumptions and behavior of the model yielded three primary implications for a theory of location learning: (1) locations are encoded as a by-product of attention, (2) once encoded in memory, location knowledge is subject to the same mechanisms as other declarative knowledge, such as associative learning and decay, such that, (3) the ability to retrieve location knowledge, like other knowledge (e.g., a phone number), requires repetition, practice, or explicit rehearsal.

The empirical results, taken together with inferences drawn from the behavior of the model, demonstrated that location learning is not only pervasive, but also subject to the cost structure of the interface. As the cost of relying on labels to locate the currently needed interface object (search cost) increased, so did the rate of location learning and reliance on location knowledge. Likewise, as the cost of relying on an object's label to evaluate whether that object is indeed the one currently needed (evaluation cost) increased, so did the reliance on location knowledge. Consistent with a rational analysis perspective, participants came to learn and rely on location more quickly when the interface provided them with no less-effortful alternative.

INTRODUCTION AND OVERVIEW

In graphical user interfaces (GUIs), the location of a given interface object, such as a button or menu item, must be determined before the object can be acted on (i.e., user must move the cursor to the location and then click). As Jones & Dumais (1986) put it "It is not enough to know what we are looking for; we must also know where to look for it" (p. 43). Knowledge of an object's location on screen is a necessary and sufficient cue for task performance - the user merely needs to move visual attention and the cursor to the location and then perform the appropriate action. Thus, knowledge of the desired object's location obviates or at least restricts the scope of the visual search undertaken to locate that object.

The restriction of the visual search space, in turn, has the potential to significantly improve task performance by reducing the time required to conduct the search, especially in visually cluttered interfaces. This is particularly relevant in light of ever-increasing functionality and concomitant visual complexity in software. As an example, Microsoft® Word 98 contains 11 menu items on the menu bar and defaults to approximately30 buttons on two toolbars, with more toolbars and buttons available as the user changes modes. Thus, a top level search (i.e., not including the items contained in the menus) would require the user to consider as many as 41, and an average of about 20, interface objects to locate the one currently needed. Even knowledge of the currently needed object's approximate location can significantly reduce the number of objects to be considered, and thus, the time required to conduct the search.

Although an object's location may be a necessary and sufficient cue that can improve performance, it is only one of a number of cues available in a typical interface. Each of these cues has strategies and costs associated with it. For example, each of the buttons on the Word toolbar has an icon. If the icon is representative of the button's function, it is possible to quickly determine whether the button will satisfy the user's current goal. If the button's function is not clear from its icon, at the cost of a mouse movement and a one second wait, a text description of the button's function (called a ToolTip) can be accessed. Because of the additional time required to access a ToolTip, the cost of a strategy that entails waiting for ToolTips can be said to be higher than one which entails simply evaluating the icon with regard to the current goal.

Rational analysis, a theoretical perspective put forth by Anderson (1991, 1993; Anderson & Lebiere, 1998), assumes that human behavior is an optimal response to the structure of the environment, such that humans tend to rely on cues and strategies that maximize the difference between cost of mental effort and expected gain in achieving goals. Taking this perspective, the cost structure of the interface (i.e., the cost associated with the available cues and strategies) has the potential to significantly impact learning and reliance on location knowledge, such that users will rely on location only to the extent that the interface provides them with no lower-cost alternative. Research demonstrating that users eventually learn interface object locations with experience, however, suggests that location learning is pervasive and may occur regardless of interface cost.

Drawing from the location learning literature and the rational analysis perspective described above, the primary goal of this dissertation is to explicate the interaction

between location learning and interface cost, and in doing so, provide a rationality-guided account of location learning in interaction with a graphical user interface (GUI). The project proceeds in two phases, an empirical phase, comprised of two experiments, and an analytical phase, in which the data from the experiments are used to constrain the behavior of a computational cognitive model that acquires and uses location knowledge. The model, by virtue of being built within the ACT-R/PM simulation environment (Byrne & Anderson, 1998), inherits components of rational analysis embedded in the underlying ACT-R cognitive architecture.

An understanding of the manner and conditions under which location information is learned as users gain experience is critical in understanding how users interact with such systems in general and can shed theoretical light on location-related HCI design guidelines. In particular, such an explanation could provide a theoretical basis for the well-known positional constancy guideline, which recommends that the location of information on screen remain constant.

The paper begins with a brief review of theoretical and applied research on memory for object locations. In particular, the section focuses on the extent to which locations are learned in various tasks and summarizes current research on location memory. This is followed by a discussion of the extent to which humans interact with their environment in a rational manner, and the development of an account of how this may affect location learning on various tasks. The hypotheses generated by this account are then introduced. The research then proceeds in the two, empirical and analytical, phases described above. The theoretical and applied implications for this work (both the empirical results and model) are then derived and discussed.

THEORETICAL BACKGROUND

Location Memory Research

Location Learning in HCI

Studies in which the locations of interface items are held constant over trials provide clear evidence that participants eventually learn the locations of those items. Most of these studies employ a search and select paradigm in which participants are given a target item and are required to find that item in some arrangement of candidate items. In some studies, the target item is a word and the candidate objects are also words arranged in various orders, e.g., alphabetically, categorized, or random (Card, 1984; Mehlenbacher, Duffy, & Palmer, 1989; Somberg, 1987; Vandierendock, Hoe, & Soete, 1988). In others, participants are given descriptions of commands and are required to select the icon or menu item associated with the command (Blankenberger & Hahn, 1991; Kaptelinin, 1993; Moyes, 1994; Moyes, 1995).

As an example, Somberg (1987) had subjects search for target words in four different orderings of menu items: alphabetic, probability of use, random, and positionally constant. The results indicated that only the subjects in the positionally constant condition improved performance over the 492 trials (and had still not reached asymptote). Further, the positionally constant group, which started out faster than the random group but slightly slower than the alphabetic and probability of use groups, was significantly faster than all of the other groups after 246 trials.

Somberg's (1987) results are consistent with those of the other research cited above; when the locations of items remain constant, performance improves over trials as participants reduce the scope of their visual search. In some of these studies, the scope of search is sufficiently reduced in later trials to eradicate performance differences between menu organization schemes and icon representativeness which existed in early trials (Blankenberger & Hahn, 1991; Card, 1984; Mehlenbacher et al., 1989; Moyes, 1995).

Studies including conditions in which the positions of items are not constant, i.e., randomized between trials, show either no improvement over trials (Kaptelinin, 1993; Somberg, 1987), or only modest improvement (Blankenberger & Hahn, 1991). In this latter case, the improvement can be attributed to continued learning of the mapping between the command used as a cue and its associated icon.

Interview-based studies of users' everyday interaction with GUIs also provide supporting evidence for a reliance on location information. Barreau and Nardi (1995) synthesized studies of Macintosh and PC users and found frequent use of desktop locations (the operating system's desktop) for organizing and reminding. A large proportion of users, for example, would place frequently used, current or otherwise significant files in specific locations on the desktop. They also found an "overwhelming" preference for location-based file searches (e.g., starting at the hard drive and drilling through the file hierarchy rather than initiating a text-based search). Kaptelinin (1996) reported similar findings.

Location Memory in Problem Solving and Reading

There are studies in the areas of problem solving and reading which also provide supporting evidence for the learning and use of location information in task performance. A study of equation problem solving by Anderson, Matessa and Douglass (1995), showed that participant performance improvement over trials was largely attributable to a decreased number of eye fixations on the equation components (i.e., numbers and operators) rather than decreased duration of the fixations. This led them to conclude that

> ...we have discovered that an important component of the learning (indeed the majority of the time savings) that is going on in this experiment is due to improved strategies for scanning the equation...This research indicates that an important component of skill development is learning where critical information is to be found in the visual interface (p. 64)

In a study on memory for words and their locations in prose, Lovelace and Southall (1983), found that providing participants with the location of a particular word as a retrieval cue significantly improved participant's recall of that word. An extreme form of this ability is reported in Stratton's (1917) description of the Shass Pollak, memory experts who are reportedly able to recall any word from the 12-volume Talmud given a page number and a location on the page. A closely related form of this ability underlies the well-known Method of Loci memory strategy, which entails placing objects to be remembered in various locations and then later using that location as a cue for recall.

The Cost of Location Learning

There is a literature to suggest that the cost of encoding object locations, a prerequisite for learning them, may be very low. In a 1979 paper, Hasher and Zacks claimed that the encoding of spatial information is an automatic process. By virtue of being automatic, the process of encoding spatial location encoding is purported to be unaffected by intent, task demands, age effects, practice, strategy manipulations or individual differences. Most of the subsequent research testing these claims uses a single-trial spatial memory paradigm in which participants study a matrix of objects for some short period of time, and then are required to either recall or recognize the objects and their locations. This is done under conditions designed to test various of the six criteria listed above.

The results from this research are mixed, with some researchers claiming that spatial encoding is automatic (Andrade & Meudell, 1993; Postma & DeHaan, 1996) whereas others claim it is not (Naveh-Benjamin, 1987; Naveh-Benjamin, 1988). The primary point of contention stems from a methodological flaw Andrade and Meudell (1993) claim exists in the Naveh-Benjamin (1987, 1988) experiments; namely that Naveh-Benjamin's test of spatial memory improperly included location memory scores for objects in the matrix that participants may not have remembered seeing. To correct the flaw, Andrade and Meudell measured contingent spatial memory (i.e., memory only for correctly recognized objects) and, consistent with Hasher and Zacks' (1979) assertion of automaticity, found no effect of task demands or intent on spatial memory.

In more recent research, Postma and DeHaan (1996) set out to evaluate the criterion that spatial encoding would not be disrupted by other task demands. They used

a task similar to the single-trial memory paradigm described above, having participants perform the task under varying levels of difficulty (number of items in the matrix) and under a dual task condition (using articulatory suppression). Based on the results of a series of experiments, the authors posited the existence of two separate processes underlying object location memory. The first process, which was demonstrated to be largely unaffected by articulatory suppression and task difficulty, is responsible for encoding the positions per se (i.e., that there was an object at a given location). The second process, which was shown to suffer under articulatory suppression and more difficult task requirements, is responsible for associating a particular object to a particular location (i.e., what was where). Thus, Postma and DeHaan made a distinction between encoding a location, which is automatic, and associating the location with an object so that may be retrieved later, which is not automatic.

Such a distinction is quite compatible with the association-based theory of memory in ACT-R/PM (Anderson & Lebiere, 1998; Byrne & Anderson, 1998). In the runnable implementation of ACT-R/PM, the locations of objects are encoded as a by-product of conducting a visual search. By virtue of having been encoded, there is a memory representation of the location which is potentially retrievable. In order to be retrieved, the representation's activation level, which can be thought of as the representation's utility given its history of use and the current retrieval cues, must exceed some threshold value. Thus, the fact that a location has been encoded provides no guarantee that it can be retrieved for later use. As a particular location is re-encoded or retrieved, its activation level, and concomitantly its probability of retrieval, increases, leading to the prediction that locations searched repeatedly can eventually come to be

retrieved and used in the context of task performance. Thus, this explanation accounts for the relatively modest incidental location memory recall in the short term (Andrade & Meudell, 1993; Naveh-Benjamin, 1987), and the finding that people do eventually seem to learn the locations of objects in the long term (e.g., Anderson et al., 1995; Blankenberger & Hahn, 1991; Somberg, 1987).

Summary

This section has summarized research from the area of HCI, problem solving, document layout and cognitive psychology on learning and reliance on object location. These studies provided evidence that people eventually learn the spatial locations of items in the world. This learning was shown to occur not only in tasks where learning locations could potentially improve performance, such as menu selection and equation problem solving, but also in tasks where no discernable benefit existed, such as reading. This suggests that location learning is quite pervasive. Although there remains some controversy as to the extent to which locations are encoded in memory automatically, i.e., without explicit effort or intent, the research suggests that locations may be learned to a modest degree as a by-product of interaction. An account of these findings was then briefly sketched in the context of ACT-R/PM.

Effort, Strategy and Rationality

Humans maintain what might be termed an *inter-formative* relationship with their environment. On the one hand, we have to exist in our environment so we shape it to suit our needs. On the other, to the extent that our behavior is constrained by and adapted to the structure of our environment, it shapes us as well. Determining whether

to change the environment or adapt to it is partially a function of the effort involved in doing one versus the other. In most everyday tasks, such as interacting with a computer, the effort required to substantially change the environment is well beyond what we are willing or able to expend (though with less effort we can write macros, modify menus, etc.), so to achieve our goals we are usually forced to adapt.

The manner in which we adapt can also be viewed as being a function of effort. Anderson has taken this perspective (1991, 1993; Anderson & Lebiere, 1998), advancing a rational theory of cognition and an associated methodology called rational analysis. Underlying both of these is the assumption that behavior is somehow optimized to the structure of the environment. Rational analysis has been applied to various aspects of human behavior such as memory categorization, causal inference and problem solving. The assumptions derived from these analyses are core components of the ACT-R cognitive architecture. The ACT-R architecture has been implemented in Common Lisp as a production system-based simulation environment, also called ACT-R. A recent addition to ACT-R is the perceptual and motor component ACT-R/PM (Byrne & Anderson, 1998), which includes a theory of visual attention and also allows explicit modeling of interaction with the external environment.

Implied by the ACT-R/PM architecture is the prediction that people will tend to behave in a manner that maximizes the difference between the cost of mental effort and expected gain. It turns out that, all other things being equal, the lowest effort strategies tend to be those that entail accessing perceptually-available information rather than those that require learning and recalling the required knowledge (Norman, 1988; Zhang & Norman, 1994). An example of this is reliance on address books or phone presets

instead of memorization of phone numbers. In this case, the individual relies on perceptual and motor operators to acquire the necessary information instead of cognitive ones.

There is ample empirical support for the idea that humans tend to choose least effort strategies. This support comes from performance-based studies, where the performance and strategies used by participants are observed and analyzed as a function of task manipulations, as well as knowledge-based studies, where participants are given recall or recognition tests to determine the extent of their knowledge about often-used objects.

Ballard, Hayhoe and Pelz (1995) found that in lieu of encoding the required information at the beginning of their task and retaining it throughout, participants instead relied on a perceptually driven strategy to pick up information as needed. Interestingly, when Ballard et al. increased the cost of information acquisition (and thus, the cost of the perceptual strategy), the frequency of this strategy decreased sharply in favor of a more memory intensive one. Lohse and Johnson (1996) also found differences in strategy as a function of the cost of acquiring information. Specifically, they found that increasing the cost of information acquisition resulted in more frugal and systematic visual search and increased dependence on working memory.

The cost of performing actions in a given task also affects the strategies chosen by participants. O'Hara and Payne (1998) found that increasing the operator implementation cost in solving the 8-puzzle (e.g., increasing the number or keypresses required to perform a given action) increased the tendency to plan problem solving actions, as opposed to the adoption of an opportunistic and perceptually-driven, hill-

climbing strategy. Gray (in press) has found that errorful and error-free performance of an interactive programming task (programming a VCR) are very well characterized in terms of the constraints imposed by the device interface and what he terms "display-based difference-reduction". Display-based difference-reduction refers to a dependence on the device interface for information on the status of the task performance, rather than maintaining that information in memory.

Zhang & Norman (1994) conducted a series of studies using isomorphs of the Tower of Hanoi and found that embedding the rules of the problem in perceptually and culturally available constraints in the task environment led to fewer errors and faster performance overall than when those rules had to be represented internally (i.e., memorized). Other research using Tower of Hanoi isomorphs has likewise revealed the importance of external problem representation in making problem solving easier (Kotovsky, Hayes, & Simon, 1985). Such a problem representation enabled problem solving to follow a perceptual strategy in which each move was cued by the perceptual states of the puzzle, rather than one involving computationally intensive planning (Simon, 1975).

If people tend to leave knowledge in the world to be gathered as needed during task performance rather than internalizing it, memory for those aspects of the world should be quite poor. There have been a number of studies that have indeed shown this to be the case. Memory for details of pennies (Nickerson & Adams, 1979), numbers and letters on telephone dials (Morton, 1967), and graphical user interfaces (Mayes, Draper, McGregor, & Oatley, 1988; Payne, 1991; Smelcer & Walker, 1993) has been shown to be remarkably poor, even for individuals with a fair amount of experience with the items.

Summary

The picture painted by this research is one in which humans are sensitive to task demands, generally taking the path of least effort. The path of least effort is often one that relies on information available in the external environment and perceptual rather than memory and cognitive processes. Dependence on cognitive operators (e.g., memory retrieval and planning) is generally minimized unless the effort required by such a dependence is less than the effort required by the alternative perceptual and motor operators.

Rationality and Display-Based Interaction

Given the critical constraining role of the environment suggested by these data, various researchers have advocated different means of incorporating features of the environment into efforts aimed at understanding or predicting behavior: a task analysis that seeks to identify the paths of rational behavior through a given task (Card, Moran, & Newell, 1983); representational analysis, where the task is broken down into its internal and external representations (Zhang & Norman, 1994); restriction of the exploration of cognition and behavior to the specific situations in which they occur (Suchman, 1987); and the development of a cognitive architecture containing the guiding principle that cognition is optimized to the statistical structure of the environment (Anderson, 1990; Anderson, 1993; Anderson & Lebiere, 1998).

A significant portion of the work prompted by this research is in the area of human-computer interaction (HCI), particularly in modeling user interaction with graphical user interfaces. The primary interpretation of this work is that users interact

with interfaces in a recognition-driven manner, such that the label of the appropriate item is recognized in the course of task performance, as opposed to being recalled prior to acting on the display. This interpretation is at the core of a number of models of display-based interaction, including D-TAG, Howes and Payne's (1990) display-based extension to Task Action Grammars (Payne & Green, 1986), Howes' (1994) Ayn model of learning menus by exploration, and Howes and Young's (1996) TAL model that learns mappings between the tasks and actions required by the interface and can perform without requiring recall of labels.

Display-based strategies are also implicated in studies and models of exploratory learning of graphical user interfaces. Most of the empirical basis for this work comes from Franzke (1994), who studied the effects of type of interface action, number of candidate objects on screen, and the quality of object label on learning by first-time users of graphing software. The results showed a heavy reliance on searching or scanning of object labels to match task goals. Several computational cognitive models have implemented this strategy including the IDXL model from Rieman, Young & Howes (1996) and the LICAI model from Kitajima and Polson (1997).

This work in display-based HCI, though capturing critical aspects of interaction, focuses primarily on the semantic properties of the display, in particular, the labels. There are a multitude of other attributes of objects on display screens which could also serve to guide and constrain performance including: the *non-textual* features of the object's label, such as an icon or a formatted label, the *size* of the object, the *location* of the object, the *state* of the object (e.g., enabled, disabled, highlighted, etc.), the object's *visible contents*, such as the items in an opened menu or in a list box, and the

object *type* (e.g., button, menu, check box, etc.). Given the quintessential role spatial location plays in GUIs, it is somewhat surprising that the relationship between display-based interaction and location learning has not yet received much attention in the literature.

Rationality, Display-Based Interaction, and Location Learning

An object's location is typically not a meaningful indicator of the underlying function or purpose of that object. Rather, its label is typically far more salient and useful (Jones & Dumais, 1986). However, an object's label can vary widely in the extent to which it represents the underlying function or meaning of the object. In interfaces with good, representative labels, even novice users can recognize an object as being the correct one because the label coincides with their existing knowledge and/or current goal (Franzke, 1994; Franzke, 1995; Jones & Dumais, 1986; Polson & Lewis, 1990). Such users can rely on the close association between the label and the function of the object from very early on in their experience with the interface. In cases where a label is poor, however, users would be forced to learn the association between the label and function. Such a case requires incurring the cost of learning to associate two previously unassociated or weakly associated items.

The cost of strengthening this association may meet or even exceed the cost of learning to associate the function of the object with its location, also an initially weak association. This suggests that users of poorly labeled interfaces may rely more on location than on the label in the context of performance. Moyes (1994) explored this question by presenting various commands to participants (e.g., delete a document) and

having them search for and select the icon associated with the command. Half of the participants used icons which were representative of the command, and the other half used icons with no relationship to the command at all (called abstract). Following five uses of each of the 17 icons in the study, Moyes switched the positions of the icons for half of the participants, and switched the labels for the other half (e.g., representative switched to abstract and abstract switched to representative). The results showed that whereas participants in the abstract position-switch condition suffered significant performance disruption, the participants in the representative position-switch condition did not, thus supporting the hypothesis that rather than incurring the cost of associating the icon with the command and relying on icon recognition, abstract position-switch participants instead relied on icon positions.

This relationship does not seem to hold as participants become more experienced. In a follow-up study, Moyes (1995, experiment 6) increased the number of trials participants completed before the switch was imposed from 5 to 20 uses of each icon and found significant performance disruption for both the abstract position switch and representative position switch groups, suggesting that both of these groups were relying in part on location. Evidence that both groups learned icon locations also came from mouse movement data indicating the presence of pre-emptive moves toward the correct icon's location before the icon was visible. This study also provided evidence contrary to the hypothesis that both representative and abstract icon groups were relying solely on location to guide performance. This evidence comes from a condition in which the icons remained functionally in the same position but the icons were all replaced with blank outlines after 20 blocks. The results showed that participants in both

the abstract to blank and the representative to blank conditions suffered significant performance disruption. If participants knew the locations of the icons and were ignoring labels altogether before the changeover condition, then there should have been no disruption. The disruption did not last long for these groups, however, as they were able to use knowledge of location to recover to previous levels of performance after only 3 blocks. Interestingly, the position switch groups were unable to recover to previous levels at all, even after 20 more blocks.

From the results of these studies emerges an apparently complex sketch of the interaction between label meaningfulness, location learning and experience. Early on in performance, if an object's label is meaningful, participants do not seem to rely on the object's location. Over trials, however, these participants acquire the ability to use location knowledge during task performance, but still seem to rely in part on the label. If the labels are arbitrary, then participants learn location earlier on in performance than if the labels are meaningful, but still appear to rely in part on the labels after many trials.

A Rational Interpretation

A more concise account for Moyes' pattern of results can by attained by examining the relative costs of the task conditions and adopting the assumption that the participants in the experiment were always choosing the lowest effort means of completing the task. The representative icons provided a lower cost means of determining what a particular button did than the abstract icons. To the extent that the abstract icons required as much or more effort to interpret than the effort required to learn and rely on location, then participants came to rely on location as well as the icons. The representative icon condition enabled a display-based strategy for evaluation, i.e.,

relying on the inherent meaningfulness of the labels in order to determine the purpose of the button instead of expending the effort to learn and rely on location.

This begs the question of why participants in the representative condition would have learned the locations at all. The answer may lie in the previously discussed research suggesting that locations can be learned incidentally. Both the meaningful and arbitrary conditions required a deliberate search of the alternative icons on screen before the correct icon could be located. Even though icons were easier to evaluate for the representative group, this group still had to evaluate multiple alternatives if the location of the correct one was unknown. Thus, even participants in the representative group could improve performance by learning locations. The critical trade-off was between the effort required to learn locations versus the effort required to attend to and evaluate multiple icons. To the extent that locations could be learned incidentally and gradually via experience, by the later trials these participants would have eventually learned the icon locations.

The explanation above draws from both the rational analysis perspective as well as the literature on location learning discussed above. If it is the case that location learning occurs as a by-product of interaction, then the locations of all positionally constant interface objects should eventually be learned with experience. However, if participants are using least effort, display-based strategies, as the rational analysis perspective implies, the rate of learning and level of reliance on location knowledge should differ as a function of interface cost, such that the rate of learning and level of reliance will increase as the cost of relying on alternative interface cues increases.

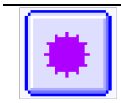
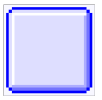
Although this theoretical account of the relationship between interface cost and location learning can explain the results of the two Moyes studies described above, a post hoc explanation is not sufficiently compelling. What is required is integrated, focused research incorporating the effects of a relatively wide range of interface costs on location learning. In the current research, interface cost is manipulated by varying the representativeness of interface object labels and location knowledge is measured at various levels of participant experience with those objects. The experimental task and specific hypotheses are outlined below.

Current Research

The theoretical account above will be explored in the context of a button-based, search and select task. In this task, participants are shown a color and twelve buttons. Each button is mapped to one of the twelve colors used in the task. They are instructed to locate and click the button associated with the color. The button locations and color to button mappings remain constant. There are four different versions of the interface for this task, each with its own set of button labels. The label sets differ in the extent to which they are representative[1] of the color applied by the button (see Table 1): the most representative labels (color-match), are blotch-shaped icons identical in color to the color applied by the button, the second most representative labels, (meaningful) are the names of the colors applied by the buttons, the arbitrary labels are icons that are not at all representative of the colors, and in the fourth version of the interface (no-label), the buttons have no labels.

[1] Label representativeness refers to the assumed strength of prior relationship between the perceptual experience of the colors used in the experiment and the label types.

Table 1. Representativeness, search cost and evaluation cost for the four label types.

Label Type		Representativeness	Search Cost	Evaluation Cost
	Color-Match	Very High	Low	Very Low
	Meaningful	High	High	Low
	Arbitrary	Low	High	Moderate
	No-Label	Low	High	High

The label types were designed so as to vary the cost of relying on the labels for two major phases of performance on this task: the search phase and the evaluation phase. The evaluation phase entails determining whether a particular button applies the desired color, and the search phase entails choosing which button to evaluate. Cost differences in the search phase (search costs) exist between the color-match condition and all other conditions. The colors on the color-match labels produce a pop-out effect, a phenomenon in which the color being sought is readily identified among a series of distractors (Triesman & Gelade, 1980; Triesman & Souther, 1985). As such, the location of the correct button can identified via automatic detection (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). The lack of a pop-out effect in the meaningful, arbitrary and no-label conditions requires that a controlled search (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977) be undertaken, i.e., deliberate eye movements be made around the interface until the required button is located. Thus, the search cost, which will be

operationalized as the number of buttons which must be evaluated before locating the correct one, is considered to be low for the color-match labels and high for the meaningful, arbitrary and no-label labels (see Table 1).

Cost differences in the evaluation phase (evaluation costs) exist between all label types (see Table 1). Evaluation cost is operationalized as the time required to determine if the currently attended button is the one currently needed. The color-match labels simply require comparing the color of the label to the rectangle color and thus incur a very low evaluation cost. The meaningful labels also only require a comparison, but because the text label must first be read before it can be compared to the name of rectangle color, the evaluation cost is slightly higher. The arbitrary labels provide no clue as to the colors associated with the buttons, but still serve to uniquely identify the buttons, thus, the arbitrary labels are considered to incur a moderate evaluation cost. If the buttons have no labels, there is a high evaluation cost.

Based on the theoretical account and the structure of the task and label types described above, there are two parallel hypotheses, one for each of the two phases of task performance.

Search Cost Hypothesis

> There is a positive relationship between search cost and reliance on location knowledge such that users of higher search cost interfaces will come to rely on location knowledge in the search phase more-so than users of low search cost interfaces.

To the extent that the interface requires a controlled search, as it would in the meaningful, arbitrary and no-label interfaces, object locations should be learned faster

than in an interface where search involves automatic detection, as it would in a color-match interface. The pop-out effect available in the color-match interface provides a low cost, efficient means of locating the currently needed button, so a rational participant should come to rely on this perceptually available cue in lieu of location knowledge. In the absence of such a cue, the only way to avoid having to conduct a higher cost, exhaustive controlled search is to learn and rely on location knowledge.

Evaluation Cost Hypothesis

> There is a positive relationship between evaluation cost and reliance on location knowledge such that users of higher evaluation cost interfaces will rely on location knowledge more in the evaluation phase than users of low evaluation cost interfaces.

As discussed above, to the extent that learning and relying on an object's label to determine if it is the object currently needed is equivalently effortful to learning and relying on its location, then a rational participant should be as likely to learn and rely on location as on the label. If the labels are representative and thus enable the use of a less effortful, display-based label-matching strategy, then a rational participant should take this available least-effort path and come to rely primarily on the label instead of the location in evaluating a given object.

Research Approach

In order to evaluate the rational account of the relationship between search cost, evaluation cost and location learning outlined above, the current research takes a two-pronged approach: the first part empirical, and the second analytical. The empirical prong

is comprised of two experiments. The first experiment employs an incidental learning paradigm in which participants perform the just-described search and select task and are unexpectedly required to rely on their location knowledge. The amount of disruption in task performance is taken as a measure of the level of reliance on location knowledge. The second experiment uses the same search and select task as Experiment I, but involves the collection of eye gaze data as a longitudinal and direct behavioral measure of location learning.

The results of these experiments are used to constrain the behavior of a computational cognitive model in the second phase of the project. The model, built using ACT-R/PM (Byrne & Anderson, 1998), is intended as a formal instantiation of the theoretical account under scrutiny, and its ability to account for the empirical data is taken as an indication of the sufficiency of the account. The model interacts with the same interfaces as the participants and, as such, its performance is constrained to coincide with key attributes of participants' behavior, from fine-grained components of interaction such as eye and mouse movements to higher order measures such as the decreased performance time resulting from location learning. The behavior and components of the model are then analyzed with regard to research on display-based interaction and current cognitive theory on location memory.

METHODS – EXPERIMENT I

Experiment I was designed to assess location knowledge under four different search and evaluation cost conditions and at two different levels of experience with interface objects. Search and evaluation cost were manipulated by varying the usefulness of the interface object labels. Location knowledge was assessed by analyzing performance when these labels were unexpectedly removed from the objects, thus forcing participants to rely solely on their knowledge of object location.

Participants

Seventy George Mason University undergraduates participated in the study for course credit. There were 18 males and 52 females approximately equally distributed within experimental groups. Participants were experienced with graphical user interfaces, reporting 532 hours of cumulative experience on average[2]. There were no differences in cumulative GUI experience between groups.

Materials

The experiment was conducted in a sound-controlled room using an Apple Power Macintosh and 17-inch color monitor. The software for presenting the experimental stimuli and collecting performance data was written in Macintosh Common Lisp.

[2] Cumulative experience was calculated by multiplying reported weekly usage (in hours) by 52 and the number of reported years of experience.

In the primary task, the software presented participants with a light gray screen that contained a centered 3 cm x 1.1 cm bordered white rectangle. After a delay of one second, a 2.8 cm x 0.9 cm colored rectangle appeared in the center of the white rectangle (see Figure 1a). This rectangle was one of 12 colors and contained seven white lower-case x's. The twelve colors were: red, blue, light blue, green, light green, tan, brown, gray, orange, yellow, pink, and purple, which were chosen to be easily nameable and discriminable. The participant then clicked on the colored rectangle to display 12 buttons arranged in a circle around it (see Figure 1b). The buttons were 1.5 cm square and had one of four label types (described below and shown in Figure 2). The participants' goal was to find, point to, and click on the button that would make the white x's the same color as the rectangle (i.e., make the rectangle appear solid). The buttons and their respective labels appeared in the same locations for a given participant, but was randomized between participants.

To find out what color was associated with a particular button, the participant could: (1) refer to the label on the button, (2) move the cursor over the button and leave it there for one second to see a ToolTip, which was a 1.5 cm x 0.8 cm bordered rectangle containing three colored x's displayed above and to the right of the button (see Figure 1c), or (3) simply move the cursor over the button and click.

When the user clicked a button, the color of the x's changed and all buttons disappeared (see Figure 1d). If the correct button was pressed, the rectangle appeared solid for 500 milliseconds and then disappeared. After a delay of 1 second, the next trial began. If the incorrect button was pressed, the computer emitted a series of 5 beeps, followed by the presentation of a dialog window informing the participant of the error.

At that point, the participant had to click a button to close the window and then would have to repeat the trial. The availability, but high time cost, of ToolTips and relatively high cost of errors were designed into the interface to facilitate and encourage the learning of button-color associations.

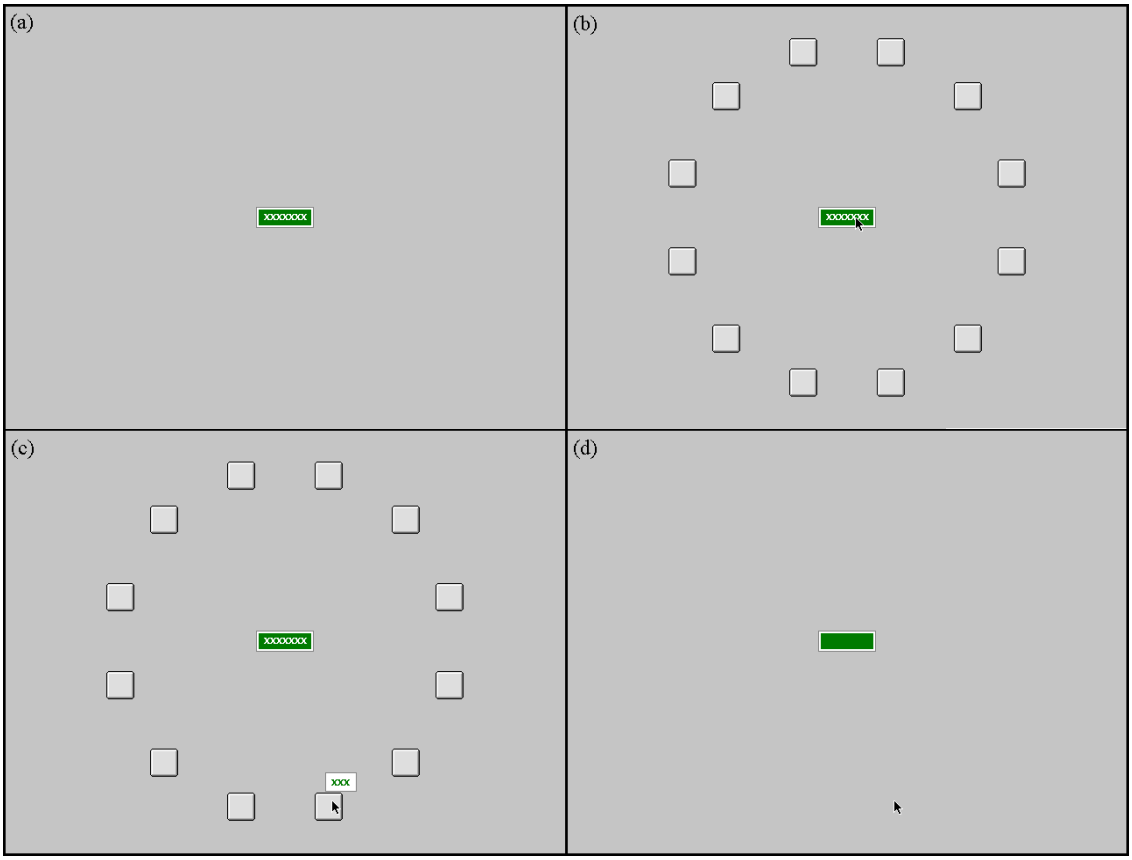


Figure 1. The time course for a trial in the experimental task.

The experimental software logged screen events such as mouse clicks on the colored rectangle and buttons, as well as accesses to ToolTips. The software also calculated performance measures, including trial time and accuracy, and wrote these data to a text file.

Design

The design was a 3 (Label) x 2 (Assessment Time) x 2 (Frequency) mixed factorial with label and assessment time as between-subjects factors and Frequency as a within-subjects factor. The fourth label condition was a control. In this condition, the buttons had no labels throughout the experiment. The design is shown in Table 2.

Table 2. The experimental design for Experiment I. Numbers in italics represent participants.

Label Type	Assessment Time	Frequency Of Use	
		High	Low
Color-Match	Early	*1…10*	*1...10*
	Late	*11...20*	*11...20*
Meaningful	Early	*21...30*	*21...30*
	Late	*31...40*	*31...40*
Arbitrary	Early	*41...50*	*41...50*
	Late	*51...60*	*51...60*
No-Label		*61...70*	*61...70*

Independent Measures

Label (Between-Subjects – 3 Levels, plus control)

As described above, the four label conditions were designed to vary search and evaluation costs. To review, search cost is operationalized as the number of buttons the participant must evaluate before locating the currently needed one, and evaluation cost is operationalized as the amount of time the participant must spend determining whether or not the currently-attended button is the one currently needed. The four

conditions are described in detail below and a summary of the assumed costs by condition is presented in Table 1.

In the color-match condition, the buttons were labeled with large colored, blotch-shaped icons. The icon color was identical to the color applied to the x's when the button was pressed (see Figure 2a). This condition was designed such that participants could search pre-attentively for the correct button by matching the color of the rectangle against the blotch on the button, thus requiring that only one button be attended per trial. As such, this condition was considered to have a low search cost. Because determining whether a given button was correct only involved making a perceptual color match between the goal color and the button's label, evaluation cost also was judged to be very low for this condition (see Table 1).

In the meaningful condition, the buttons were labeled with color names (see Figure 2b). As such, participants had to perform a controlled search of the interface to find the correct button (i.e., the correct button could not be located pre-attentively as in the color-match condition). Because this would require attending to, on average, 6 buttons per trial, this condition was considered to have a high search cost. However, because the participant needed only to read the text on the label to verify that it was the correct one, evaluation cost was considered to be low (see Table 1).

In the arbitrary condition, the buttons were labeled with icons bearing no relationship to the color applied by the button (e.g., the button that applied blue had a plane icon; see Figure 2c). The mapping from color to icon was determined randomly at the beginning of the experiment and remained constant throughout. As in the meaningful condition, participants could not pre-attentively locate the correct button, thus incurring

a high search cost. Because evaluation required learning the association between the icon and the color applied by the button, relying on the ToolTips, or learning and relying on location, this condition was considered to have a moderate evaluation cost (see Table 1).

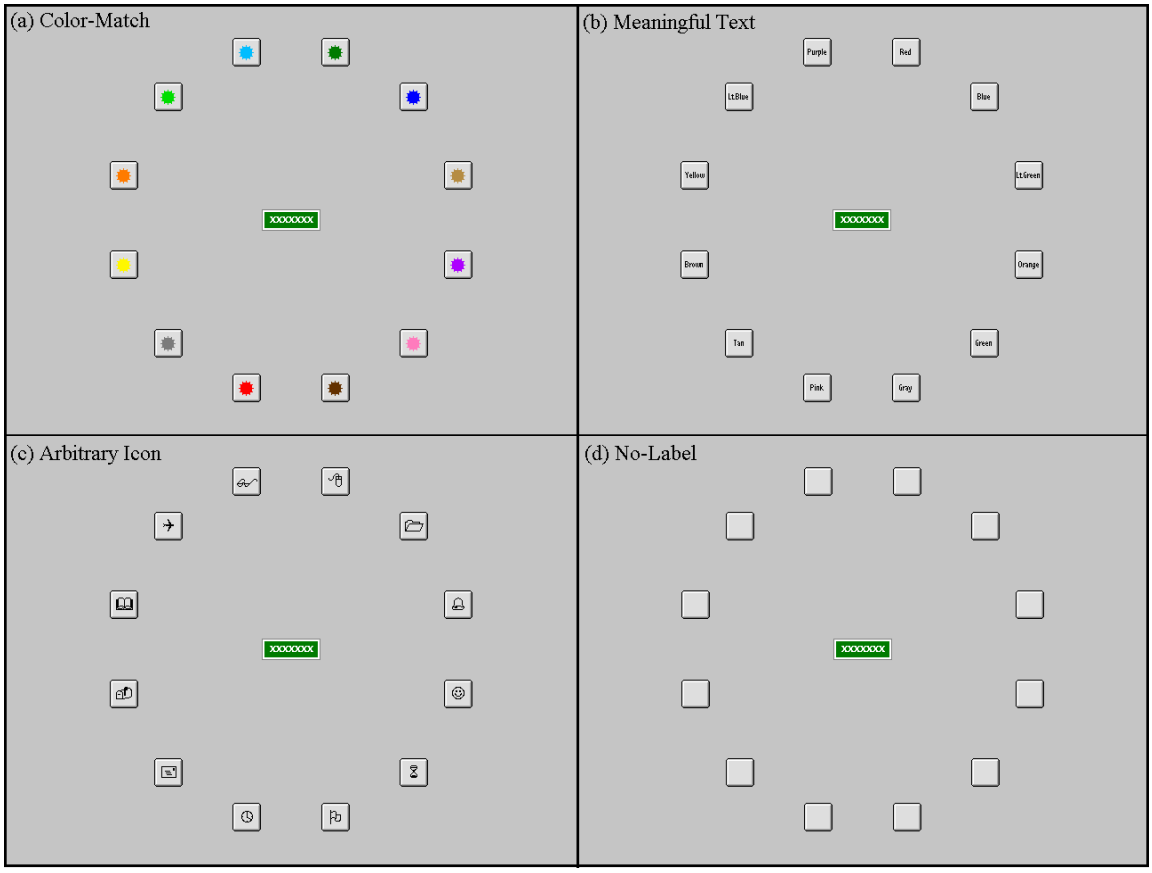


Figure 2. Screen snapshot for each of the four conditions in the experimental task.

Finally, in the No-Label condition, the buttons were all unlabeled (see Figure 2d). Again, as in the meaningful and arbitrary conditions, participants could not locate the correct button pre-attentively, thus incurring a high search cost. The absence of labels

forces participants to rely on ToolTips or location knowledge, thus entailing a high evaluation cost (see Table 1).

Assessment Time (Between-Subjects– 2 Levels, nested within Label)

For the six experimental groups, reliance on location knowledge was assessed either early (after 6 blocks) or late (after 26 blocks) in the experiment. The assessment method entailed removing the button labels to force participants to rely on their location knowledge, and measuring the effect on performance, e.g., if a participant showed no disruption in performance after the labels were removed, it would be inferred that the participant had learned the object locations. To minimize disruptive effects in performance due to surprise or confusion, participants were shown the following message during the one-minute break given just prior to the assessment block:

> "When you return to the task, the buttons will all be in the same positions as before, but they will no longer have icons on them. They will stay this way for the rest of the task. Hint-boxes will still be available."

For the meaningful group the word "text" was substituted for "icons". By giving warning just prior to removal of the label, participants had only minimal time to prepare for the change in conditions but yet were not surprised by the sudden sight of blank buttons on the screen.

Frequency (Within-Subjects –2 Levels)

In an attempt to investigate the effects of experience on location learning at a relatively fine grain, the 12 buttons used in the interface were randomly divided into two equal-sized groups at the beginning of the experiment. Six buttons were assigned to the

High Frequency group; these were needed twice per block. The six buttons in the Low frequency group were only needed once per block.

Procedure

The experimental procedure is represented schematically along with the experimental protocol in Appendix A. Participants began by completing a series of computer-presented pre-tests and training under the supervision of the experimenter. This was followed by the main experimental task (described above), a series of post-tests, and an on-line questionnaire on computer experience, which were all completed in the experimenter's absence. Finally, participants were debriefed and asked about their strategy use. The pre-tests, training and post-tests are described below; screen snap-shots of the tasks appear in Appendix B.

The first task was a color training task, which was designed to introduce the colors needed in the main experimental task and also to ensure that participants had adequate color vision. Participants were shown a colored rectangle which was one of the 12 experimental colors and were instructed to click on the name of that color from a pop-up menu that appeared to the right of the rectangle (**Error! Reference source not found.** in Appendix B). If an error was made, the correct response was provided via a dialog box. Participants were required to successfully complete at least one out of the three trials with each color before continuing the experiment.

The next task was designed to control for individual differences in mouse movement speed in later analyses. In this task, participants were instructed to move to and click on a button which changed position when clicked (**Error! Reference source**

not found. in Appendix B). The button began in the center of the display and returned to the center position on every other trial. When the button was not in the center it appeared randomly in one of the 12 positions used in the main experimental task. The button appeared five times in each of these 12 positions.

To introduce participants to the flow of the main experimental task, but not the specific stimuli used, participants next completed a training task. This training task was identical to the main experimental task, except that the colored rectangle with white x's was replaced by a 6 cm gray-patterned square with a 3 cm diameter white circle, and instead of 12 buttons arranged in a circle, there were four buttons (one for each of the four possible gray-patterns) arranged in a column to the right of the square (**Error! Reference source not found.** in Appendix B). Participants performed the task three times for each of the four patterns, and were provided guidance only if necessary.

After the training, participants completed the main experimental task under one of the seven conditions for 30 blocks of 18 trials, for a total of 540 trials. They were given three one-minute breaks, one each at the end of blocks 5, 15 and 25.

To determine the precision of the participant's knowledge of the button locations at the end of the experiment, participants were presented with a colored rectangle in the center of an otherwise blank screen and instructed to move the cursor, which dragged with it a square the same size as the buttons, to the location of the correct button (i.e., the button associated with the current rectangle color) and click to drop the square (**Error! Reference source not found.** in Appendix B). Once the square was dropped, two buttons appeared at the bottom of the screen; one button allowed participants to repeat the square placement if they were not satisfied with the current one, and the

other would advance to the next trial. Squares dropped by participants were erased when either of the two buttons were pressed. Participants completed this task for all 12 buttons.

To determine the extent to which the participants in the arbitrary group learned the association between the icons and the colors, participants in this group were shown a button with an iconic label and were required to answer two questions: (1) Was this icon used in this study? and (2) If so, which color was associated with it? Each of these questions was answered via selection from pop-up menus such that the menu for question 2 was disabled unless the answer to question 1 was affirmative (**Error! Reference source not found.** in Appendix B). There were 24 trials, including 12 with the icons used in the experiment and 12 foils. Question one required recognition of the icon, and question two required retrieval of the color associated with it.

Analyses/Predictions

As mentioned above, the primary means by which location knowledge is evaluated is via unexpectedly removing the button labels, thus forcing participants to rely on their knowledge of button locations. The level of disruption in participants' performance assumed to be inversely related to their level of location knowledge. As indicated in the hypotheses, the predictions rely on the assumption that the more representative the labels, the poorer the location knowledge. Table 3 summarizes the performance disruption predictions for Experiment I.

Table 3. Predicted levels of disruption for assessment time and label conditions when compared with baseline performance, disruption levels for other label groups, and disruption levels for early

versus late assessment times.

Assessment Time	Label Type	vs. Baseline	vs. Color-Match	vs. Meaningful	vs. Early
Early	Color-Match	>			
	Meaningful	>	<		
	Arbitrary	=	<	<	
Late	Color-Match	>			<
	Meaningful	=	<		<
	Arbitrary	=	<	=	=

In terms of the search cost hypothesis, participants in the color-match condition are expected to show not only significant performance disruption as compared to baseline, but also significantly more disruption than both the meaningful and the arbitrary groups, at both the early and late assessment times (Table 3, columns 3 and 4). Participants in the color-match condition are expected to rely on the pop-out effect in the search phase instead of a more time-consuming controlled search, and thus learn location at a slower rate. The color-match condition is expected to learn locations over trials, however, so the level of disruption is expected to be smaller at the late assessment time than at the early time (Table 3, column 6).

Participants in the meaningful group are expected to conduct a controlled search during the search phase but to rely on a display-based, label recognition strategy during the evaluation phase. As such, they are expected to show significant levels of disruption at the early assessment time, but not at the late time (Table 3, column 3). As compared to the arbitrary group, who are expected to learn locations from very early on, the meaningful group should show more disruption in the early phase, but an equivalent and non-significant level of disruption at the late assessment time (Table 3, column 5).

RESULTS - EXPERIMENT I

Performance Curves

To evaluate the assumptions about the cost differences between label groups, an analysis was performed on the nature of improvement in each of the label conditions. For several reasons, the analysis was restricted to blocks 1-10 for the no-label and late groups. First, trial times for blocks 1-30 (see Figure 3), suggested that all seven groups had not reached equivalence by block 5, the pre-assessment block for the early groups, necessitating an analysis that looked at performance beyond this point. Second, given the increases in trial time due to the removal of the labels in block 5 for the early assessment group, it would be inappropriate to perform trend analyses including trial times from blocks 6-10 for these early groups. Third, it appears from Figure 3 that within each of the label groups the early and late conditions show identical performance through block 5. This was confirmed via a 3 (Label) x 2 (Assessment Time) repeated measures ANOVA run on trial time for blocks 1-5 for all groups, which revealed a non-significant main effect of assessment time, $\underline{F}(2, 54) < 1$, $\underline{ns}$. (The probability of a Type I error was set at .05 for all analyses reported in this paper.) Thus, the analyses described below may be considered representative of *all* participants within a given label condition.

A 4 (Label) x 10 (Blocks) repeated measures ANOVA run on trial time for blocks 1-10 for the late and no-label groups yielded significant main effects of label, $\underline{F}(3,36) = 17.24$, $\underline{p} < .05$ and blocks, $\underline{F}(9, 324) = 120.95$, $\underline{p} < .05$, both of which were superceded

by a significant Label x Blocks interaction, $F(27, 324) = 24.28$, $p < .05$. This interaction was investigated in more depth via an analysis of simple main effects.

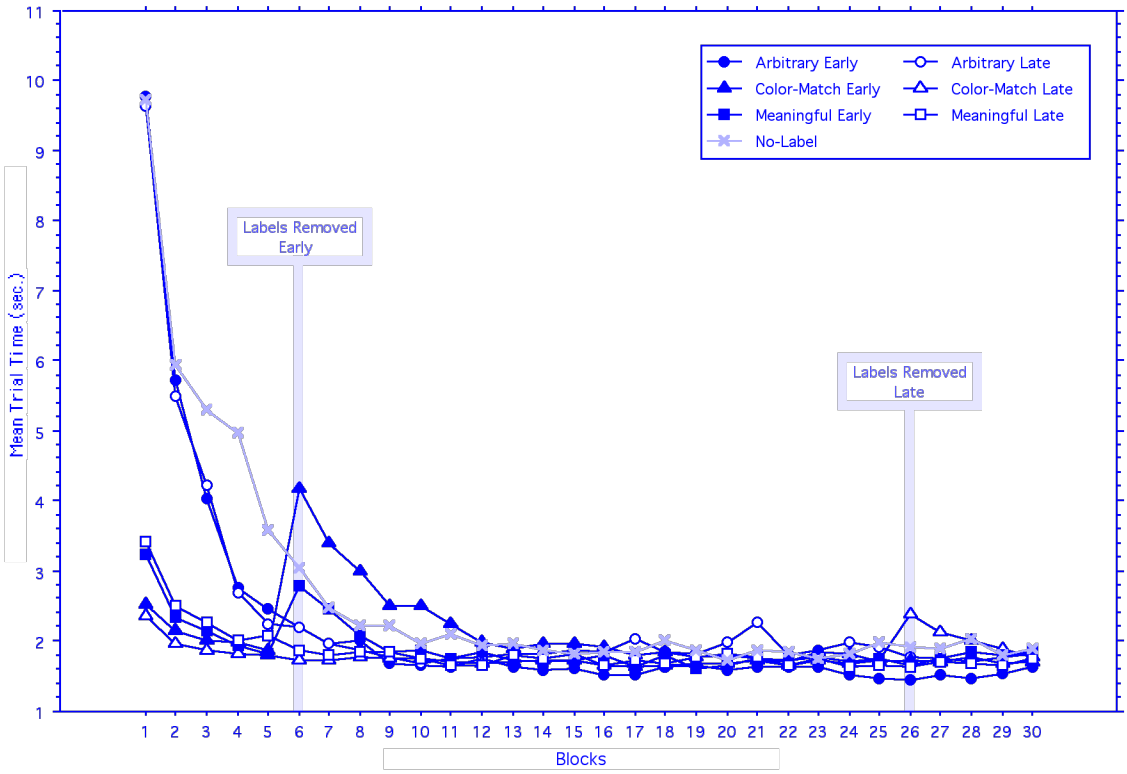


Figure 3. Trial times from Experiment I by label condition and blocks. Gray signposts denote assessment blocks.

As can be seen in Figure 4, the color-match and meaningful text groups both started off quite fast and improved only slightly as compared to the other two groups, as indicated by relatively flat performance curves. Performance improvement for the color-match group showed no significant change over blocks, $F(9, 324) < 1$, ns, suggesting that participants in this group were indeed conducting a pre-attentive search for the buttons throughout the task. In contrast, the arbitrary, meaningful, and no-label

groups all showed significant improvement over the 10 blocks, Fs(9, 324) = 99.29, 4.05, and 89.86, respectively, ps < .05.

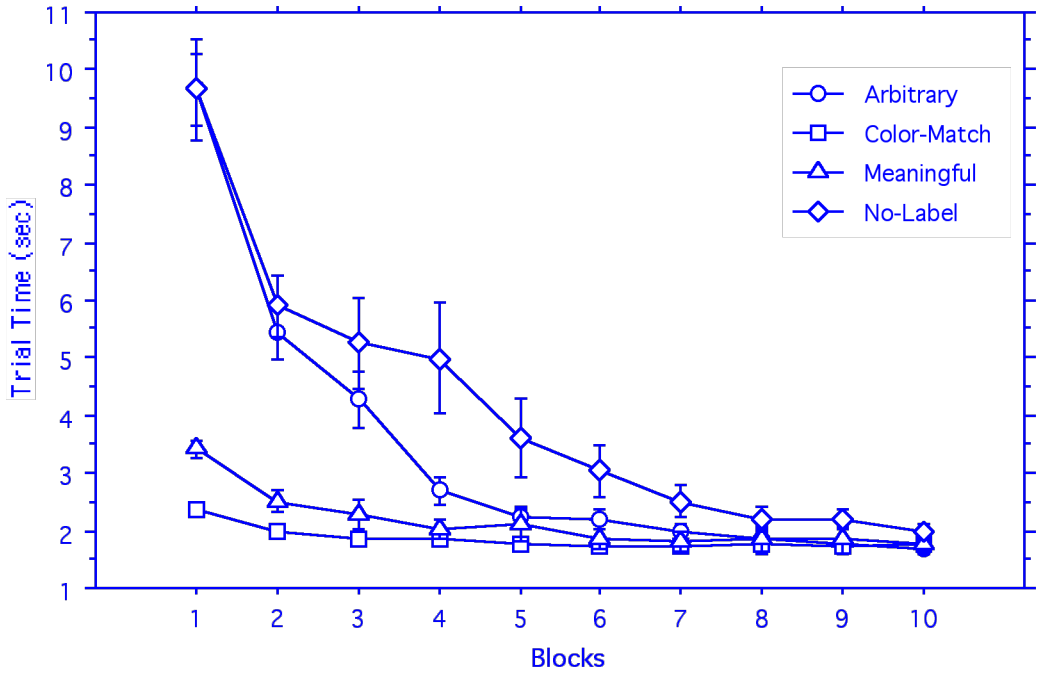


Figure 4. Trial times for the first 10 blocks for the late and no-label groups. Error bars represent standard error.

The analysis of simple main effects revealed that label groups significantly differ up through block 6, after which these differences are no longer reliable F(3, 120) < 1, ns. Including only these first 6 blocks, comparisons were run between label groups within the Label x Blocks interaction. This analysis revealed that the interface cost differences designed into the label conditions for the most part appear in the trial times. The no-label group took longer to reach asymptote than the arbitrary group, F(5, 180) = 4.83, p < .05, and the arbitrary group took longer than the meaningful text group F(5, 180) =

41.46, $p < .05$. The contrast between the meaningful text group and the color-match group was not significant, $F(5, 180) < 1$, ns.

Figure 5 reveals that nature of the curves presented in Figure 4 are consistent with the power law of learning. This law, which maintains that the acquisition of knowledge and skill can be described by a power curve, essentially means that there are initial large gains in performance followed by a more gradual rate of improvement. As shown in Figure 5, power curves provide an excellent fit to each of the four groups, ranging from an r^2 of .86 for the color-match group to .98 for the arbitrary group.

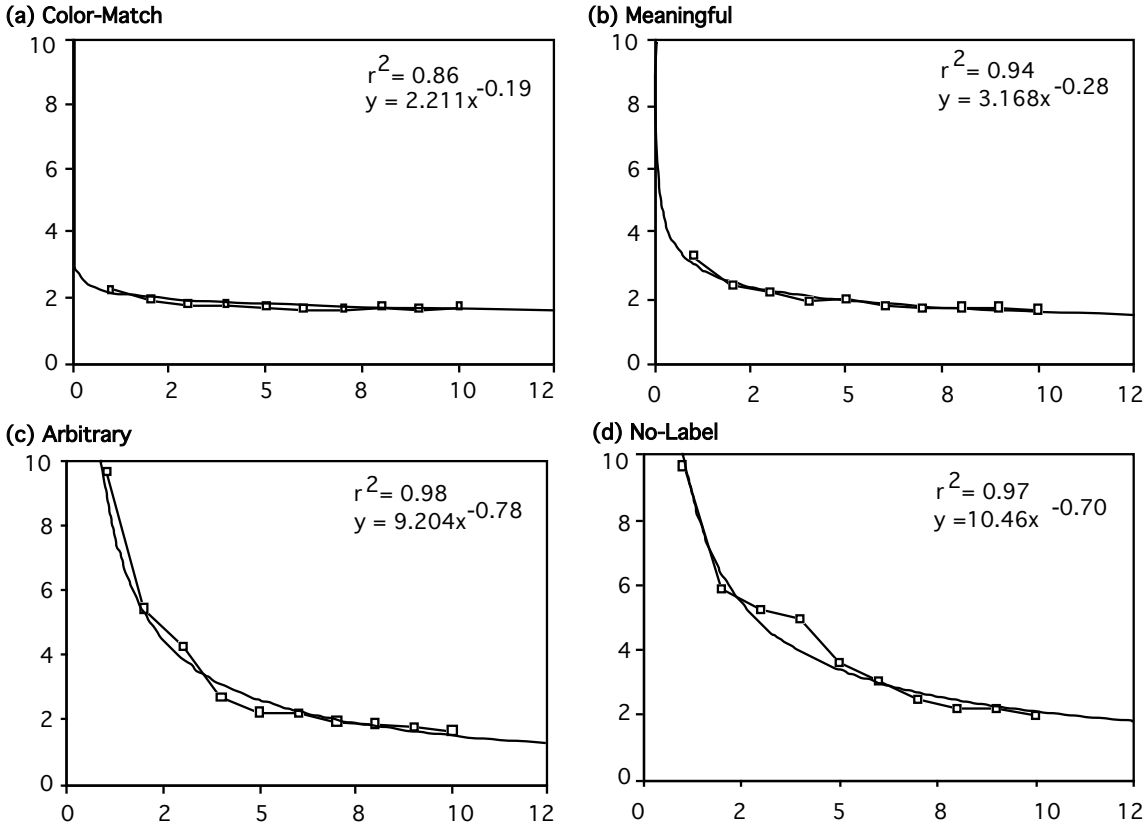


Figure 5. Power curve fits to the Experiment I trial times (in seconds). Data are from blocks 1-10 for the late and no-label groups only.

Performance Disruption

One critical assumption underlying the experimental method was that trial time and errors would increase after the labels were removed, due to participants' inability to recall correct button locations. Thus, significant disruptions in performance would be assumed to reflect poor location knowledge. As such, two disruption scores were calculated, one for accuracy and one for trial time.

Accuracy Disruption

The experimental task was designed to minimize errors by imposing a relatively high error cost. As a result, accuracy was quite high overall at 98%. The accuracy disruption score was calculated by subtracting the mean accuracy of the last use of each of the 12 buttons in the pre-assessment block (block 5 for the early groups and 25 for the late groups) from the mean accuracy of the first use of each of the 12 buttons in assessment blocks (block 6 for the early groups and 26 for the late groups). Means and standard deviations are shown in Table 4.

A 3 (Label) x 2 (Assessment Time) x 2 (Frequency) mixed ANOVA with label and assessment time as between-subjects factors and frequency as a within-subjects factor was run on accuracy disruption score. This analysis yielded a significant main effect of label $F(2, 54) = 3.45$, $p < .05$, and a significant three-way Frequency x Label x Assessment Time interaction $F(2, 54) = 3.07$, $p = .05$.

Planned comparisons (see Table 3) of the label main effect revealed that the color-match group showed a significantly larger drop in accuracy than the meaningful and arbitrary groups (Fs(1,54) = 4.11 and 6.06, respectively, ps < .05), but that there was no difference between the meaningful and arbitrary groups, $F(1, 54) < 1$, ns. Thus, the controlled-search groups retained higher accuracy than the color-match group when the labels were removed. This was also borne out in individual t-tests run to assess whether or not the level of disruption was significantly different from 0. These t-tests, shown in Table 4, revealed that although all three label groups showed significant levels of disruption in early assessment, only the color-match group showed a significant accuracy drop in late assessment. Thus, the color-match group, even after completing

450 trials, had less accurate knowledge of button locations than either the color-match or the arbitrary group.

Table 4. Means (standard deviations) and t-test for difference from zero for accuracy disruption score (in proportion correct) by label and assessment time.

	Early			Late		
Label	M (SD)	t	p	M (SD)	t	p
Color-Match	-0.10 (0.16)	-2.03*	0.04	-0.09 (0.11)	-2.54*	0.02
Meaningful	-0.07 (0.08)	-2.75*	0.01	-0.01 (0.03)	-1.00	0.17
Arbitrary	-0.04 (0.07)	-1.86*	0.05	-0.01 (0.03)	-1.00	0.17

*p<.05, df=9, one-tailed

The three-way Frequency x Label x Assessment Time interaction is due to the meaningful early group differing in accuracy as a function of button frequency. The meaningful early group shows more accuracy disruption for the low (M_{Low} = -0.13) than the high (M_{High} = 0.00) frequency buttons. There is no particular theoretical interpretation for this; it is taken as random variation in the data.

Overall, accuracy was high for all groups at 98% correct. All three label groups showed significant drops in accuracy in the early condition, but only the color-match group showed a significant drop in the late condition. Comparisons between the label groups revealed that the color-match group showed more disruption than the arbitrary or meaningful groups, but no difference between the meaningful and arbitrary groups. Thus, participants in the lowest interface cost (color-match) condition showed the most disruption.

Time Disruption

A time disruption score was calculated for each button for each participant. Calculating the score involved subtracting the trial time for the last correct use of each button in the pre-assessment block from the trial time for the first correct use of that button in the assessment block. Means and standard deviations are reported in Table 5.

Table 5. Means (standard deviations) and t-test for difference from zero for time disruption score (in seconds) by label and assessment time.

	Early			Late		
Label	M (SD)	t	p	M (SD)	t	p
Color-Match	2.97 (1.39)	6.75*	0.01	1.15 (1.47)	2.47*	0.02
Meaningful	1.14 (1.17)	3.08*	0.01	0.05 (0.42)	0.35	0.36
Arbitrary	-0.03 (0.83)	-0.10	0.46	-0.20 (0.47)	-1.32	0.11

*p<.05, df=9, one-tailed

A 3 (Label) x 2 (Assessment Time) x 2 (Frequency) mixed ANOVA with label and assessment time as a between-subjects factor and frequency as within-subjects was run on time disruption score. This analysis yielded significant main effects of both label and assessment time Fs(2, 54) = 23.11 and 14.05, respectively, ps < .05, as well as significant Label x Assessment Time, $F(2, 54) = 3.29$, $p < .05$, and Frequency x Assessment Time, $F(1, 54) = 4.53$, $p < .05$, interactions.

Planned comparisons within the Label x Assessment Time interaction (see Table 3) revealed that the color-match-early group was significantly more disrupted than the meaningful-early group, $F(1,54) = 15.87$, $p < .05$, and the meaningful-early group was,

in turn, more disrupted than the arbitrary-early group, $F(1,54) = 6.43$, $p < .05$, thus supporting the hypotheses that higher search and evaluation costs increase the acquisition of location knowledge. Also supporting the hypothesis were individual t-tests which revealed that the color-match and meaningful early groups showed significant time disruption but that the arbitrary group did not (see Table 5).

These relationships did not hold for the late assessment condition, however, in which there was a significant difference between the color-match-late and meaningful-late groups, $F(1,54) = 5.72$, $p < .05$, but no difference between the meaningful-late and arbitrary-late, $F(1,54) < 1$, ns. Unlike the other two groups, the color-match-late group was still showing significant performance disruption in late assessment (see Table 5). Thus, in contrast to the color-match condition, the controlled search groups had learned most or all of the button locations by 25 blocks into the experiment.

Investigation of the Frequency x Assessment Time interaction indicated that the level of disruption on high versus low frequency buttons was greater at late assessment than early assessment. Thus, the less experienced early assessment participants were equally unfamiliar with the locations of the high and low frequency buttons as compared to the more experienced participants in the late condition, who were disproportionately more familiar with the high frequency buttons, and thus recovered more gracefully with these buttons.

Response Categories

Central to the hypotheses being tested is the level of participants' location knowledge at various points in the study. In order to precisely explore knowledge of

location, an encoding scheme was applied to the assessment block for the label groups and the corresponding blocks (6 and 26) for the no-label group (to enable comparison of this control group to the others). For these blocks, the first use of each of the buttons was encoded into one of four mutually exclusive and exhaustive categories. Trials were encoded as Direct if the participant clicked on the correct button without accessing any tips, Verify if the participant accessed a tip on only the correct button and then clicked, Search if the participant accessed tips on more than one button but eventually clicked the correct one, and *Miss* if the participant clicked the wrong button. The distributions of categories within label condition are presented in Figure 6.

Based on these categories, a Location Knowledge Score (LKS) was calculated as the percent of directs and verifies. The assumption behind including verifies was that in these trials participants *knew* the location of the correct button, but could not evaluate the button without accessing a ToolTip. Thus, LKS may be considered a sensitive and somewhat liberal measure of location knowledge. A 3 (Label) x 2 (Assessment Time) x 2 (Frequency) mixed ANOVA with label and assessment time as between-subjects factors and frequency as within-subjects was run on LKS. This analysis yielded significant main effects of label, $F(3,72) = 13.74$, $p < .05$, and assessment time, $F(1,72) = 30.08$, $p < .05$, but no significant Label x Assessment Time interaction, $F(3,72) < 1$, ns.

The significant main effect of assessment time indicates that locations were indeed learned with experience. Not surprisingly, the groups overall knew a significantly larger percentage of locations after 26 blocks of trials than they did after only 6 blocks ($M_{early} = 74.8$, $M_{late} = 92.9$).

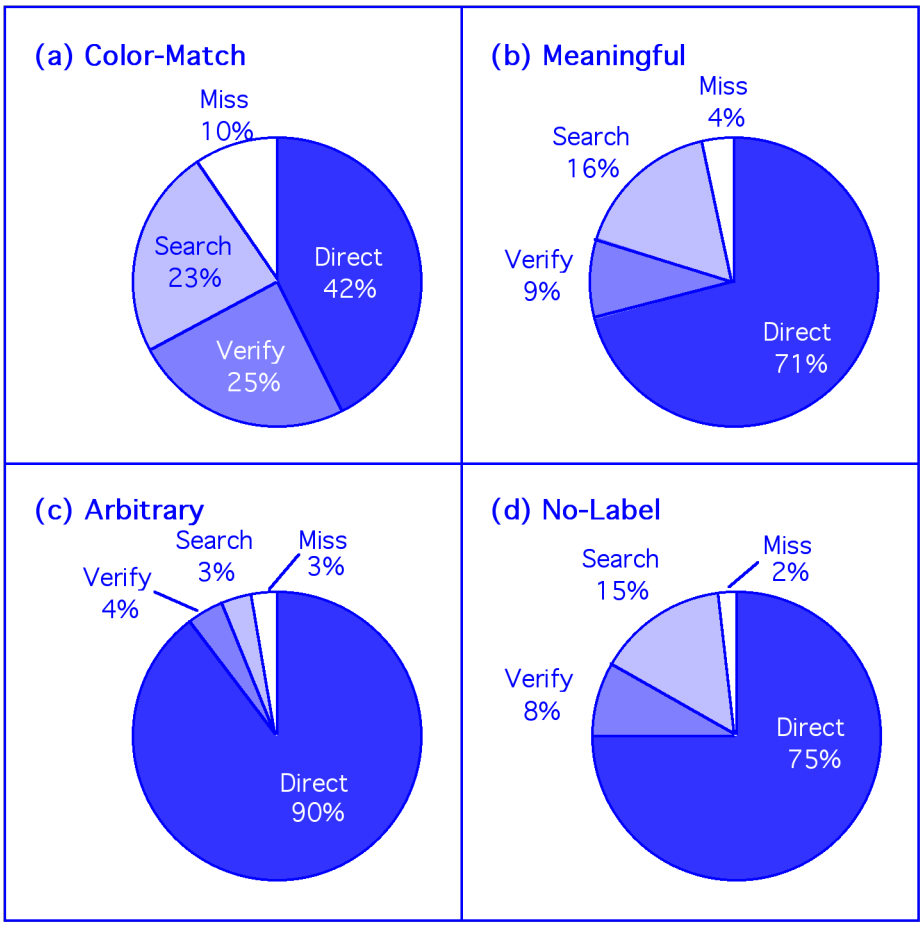


Figure 6. Distribution of response categories for post-assessment trials collapsed over early and late assessment times.

Comparisons within the label main effect revealed that the color-match group performed worse than the combination of the three controlled search groups, ($M_{color\text{-}match}$ = 65.9, M_{Others} = 89.9, $F(1,72) = 39.64$, $p < .05$), but that no differences existed between these three groups ($M_{meaningful}$ = 87.1, $M_{arbitrary}$ = 93.0, $M_{NoLabel}$ = 89.6), $Fs(1,72)$ = 1.56, .29 and .51, ns, thus providing support for the search cost hypothesis. The increased cost of searching the interface for the needed button resulted in the

controlled search groups learning more locations than the group not forced to incur this cost.

LKS measures the reliance on location knowledge in the search phase, but does not directly measure the extent to which groups were relying on this knowledge in the evaluation phase of task performance. Under the assumption that verify trials imply a weak prior reliance on location knowledge in the evaluation phase, an analysis of the frequency of these trials was undertaken. This analysis revealed a large difference in the ratio of the number of Directs to Verifies between the arbitrary (22:1), and meaningful (8:1) groups, indicating that participants in the arbitrary condition were far less likely to rely on ToolTips prior to clicking a button. Indeed, when LKS was recalculated with Verifies excluded, the arbitrary and meaningful groups were found to be significantly different ($M_{meaningful}$ = 70.9, $M_{arbitrary}$= 89.6), $F(1,72) = 5.68$, $p < .05$. This finding yields support for the evaluation cost hypothesis, which predicted that the meaningful group would rely less on location than the arbitrary group.

Proximity Analysis

The analysis above used the response categories to focus on the accuracy of the location knowledge; this analysis focused on the precision of that knowledge. For search and miss trials, the distance (in buttons) was calculated between the correct button and the first button for which a tip was accessed in a search trial or incorrectly clicked in a miss trial (see Figure 1). This analysis revealed a disproportionate number of trials in which participants accessed a tip or incorrectly clicked a button that was adjacent to the

correct button, suggesting that participants often knew the approximate locations of the correct button.

To explore this further, an attempt was made to filter out trials in which participants accessed a tip on an adjacent button by chance. As such, a trial was considered indicative of approximate location knowledge (an approximate-trial) only if it met one of two criteria: (1) it was a miss trial or (2) it was a search trial which met the additional constraint that two or fewer tips had been accessed over the course of the trial. This second constraint excluded trials in which the participant first accessed a tip on an adjacent button but then accessed tips on several other buttons before eventually locating the correct one. If the participant knew the approximate location, then there would only be a tip on the adjacent button, then either a tip access or a click on the correct one, for a total of two or fewer tips.

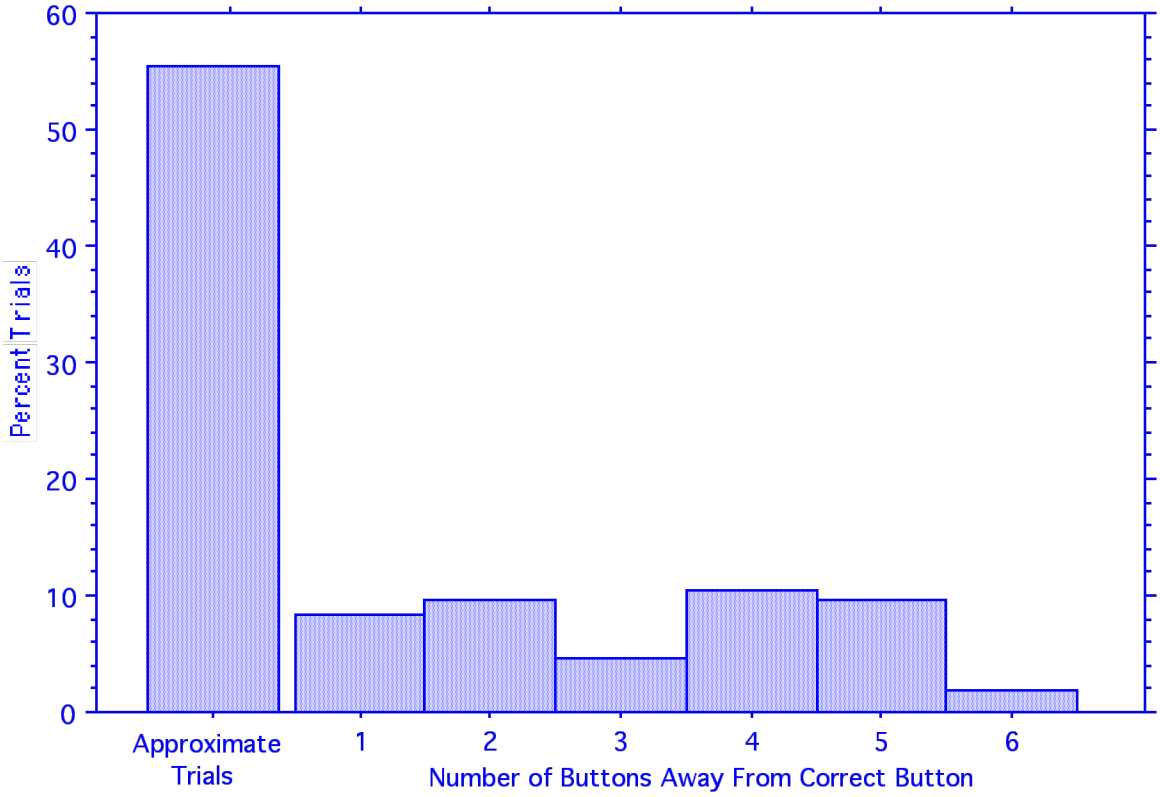


Figure 7. Distribution of errors and incorrect first guesses for post-assessment trials. The distance metric is the distance in number of buttons from the error or first guess trial to the correct button (see Figure 8). Approximate-trials are distinct from the reported 1s in that to be counted as an approximate-trial, the participant had to either make an error on an adjacent button or click the correct button directly after accessing a ToolTip on an adjacent button. Thus, 1s include trials in which more than two ToolTips were accessed; the implicit assumption being that the participant attended to an adjacent button merely due to chance.

As can be seen in Figure 7, which shows the distribution of distances collapsed over all groups, there is a disproportionate number of approximate-trials. In 55.5% of the assessment trials, when participants did not know the exact location of the needed button, they appeared to know its approximate location (plus or minus one button). Table 6 shows the mean and standard deviations of percent approximate-trials broken down by type of label and assessment time. The cell size within groups is too disproportionate and small to enable an appropriate statistical analysis (e.g., the

arbitrary-late group has an n of 1), but both within and between levels of the assessment time manipulation, the groups show a similar proportion of approximate-trials, suggesting that this phenomenon may be global and pervasive.

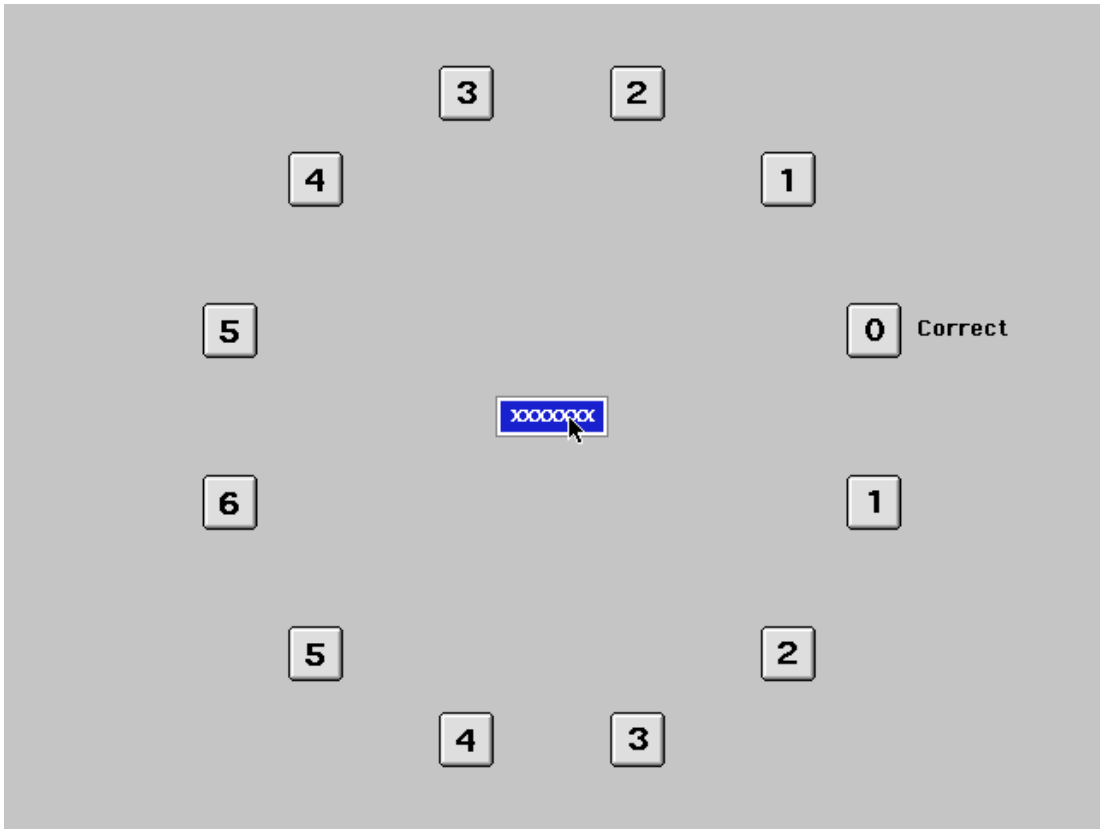


Figure 8. Example calculation of distance metric used in proximity analysis. Distance values ranged from 0 to 6 and are in units of buttons.

Table 6. Means and standard deviations for proportion of approximate-trials by label and assessment time

	Early		Late	
Label	M	SD	M	SD

Color-Match	0.40	0.50	0.64	0.46
Meaningful	0.65	0.45	0.80	0.55
Arbitrary	0.56	0.51	1.00	
No-Label	0.64	0.51	0.67	0.58

Icon Memory Test

Participants in the arbitrary condition were given a series of tests at the end of the study to determine the extent to which they learned the association between the icons and the colors (**Error! Reference source not found.** in Appendix B). The icon recognition test required participants to identify the icons used in the experiment among a series of foils. Results showed recognition was equally high for both early and late groups, with means of 94% and 93% correct respectively, $F(1, 18) < 1$, ns.

The results of the association test, which required identifying the color associated with a particular icon, are very similar to the pattern of results of the recognition test except that performance was not as high. The difference between the early and late groups was not significant (M_{late} = 88%, M_{early} = 73%), $F(1, 18) = 2.20$, ns.

The lack of differences between groups is quite surprising given that the late group used buttons with the icons on them for 5 times as many blocks as the early group. What is especially surprising, is the relatively high score for the early group on the association test (73%), given that these participants had only seen the icons for 90 trials, and had not seen them for the last 450 trials. It is possible that participants in the early group had, at least to some extent, learned the association between the icons and

colors early and continued to rely on this knowledge even after the labels were no longer displayed on the buttons. Implied in the only marginally better performance of the late group is that this group was relying on knowledge of the color-icon associations about as much as the early group, who no longer had the icons visually available.

Location Memory Test

In the location memory test, participants were shown a colored rectangle and instructed to place a button-sized square as precisely as possible in the location of the button associated with the current color (see **Error! Reference source not found.** in Appendix B). The performance measure for this test was the mean absolute euclidean distance in pixels from the center of each square placement to its respective correct center location. The means and standard deviations are presented in Table 7.

Table 7. Means, standard deviations and sample sizes for accuracy scores on 1ocation memory test. The scores are the average euclidean distance in pixels from the guessed locations to their correct respective button location.

Label Condition	M	SD	n
Color-Match-Early	25.01	4.48	10
Color-Match-Late	37.83	16.40	9
Meaningful-Early	30.29	9.01	10
Meaningful-Late	32.35	11.19	10
Arbitrary-Early	25.63	9.54	10
Arbitrary-Late	26.47	9.94	10
No-Label	29.56	12.69	9

Overall	29.59

Data from two participants were excluded from these analyses as it was apparent from a visual inspection of the pattern of their square placement that instructions for completing the task were not followed (e.g., buttons were placed in the center rectangle or over the control buttons in the lower right hand corner of the screen). One participant was from the color-match-late group and the other was from the no-label group.

The scores were subjected to a one-way ANOVA with group as the between-subjects factor, which revealed no significant differences between the groups, $F(6, 61) = 1.63$, ns. Thus, by the end of the experiment all groups had apparently learned the button locations to an equivalent extent. It is interesting to note that the overall mean score for the groups was 29.6 pixels, which, given that each of the buttons used in the experiment was 32 pixels square, means that, on average, participants were able to place buttons in locations that overlapped the 32 pixel boundary surrounding the center of their correct locations. This shows that location memory, on average, was quite accurate.

DISCUSSION - EXPERIMENT I

The performance curve analysis (see Figure 4) demonstrated that the groups in the four label conditions performed in a manner consistent with the anticipated search and evaluation costs outlined in Table 1. The curve for the lowest cost, color-match, condition (low search and very low evaluation cost) was quite flat as compared to the other conditions. Indeed, the analyses revealed that this group, unlike any other group, showed no significant improvement. The meaningful text group, designed to have the second-lowest cost overall (high search and low evaluation cost), did show significant improvement over blocks, but still showed a relatively flat performance curve as compared to the arbitrary and no-label conditions. The meaningful group started off almost three times faster than the arbitrary and no-label groups, which both showed large and significant improvement in performance over trials. The arbitrary and no-label groups did not show identical improvements in performance, however, supporting the distinction between the second-highest cost, arbitrary (high search and moderate evaluation cost) condition and the highest cost, no-label (high search and high evaluation cost) condition.

Location Learning

Consistent with the research discussed in the literature review, the results of Experiment I provided ample evidence that all groups learned button locations. The time disruption analysis revealed that participants showed a significantly larger disruption in

performance in the early assessment condition than in the late assessment condition, indicating that participants in the late groups had more accurate location knowledge. Similarly, the analysis of location knowledge score revealed that participants overall knew a significantly larger percentage of locations in the late assessment condition than in the early assessment condition.

The results of the location memory test analysis also provide strong evidence for location learning. By the end of the experiment, participants were able to reconstruct the original locations of the buttons with a high degree of accuracy, on average overlapping the correct location with their placement.

The proximity analysis suggests that location learning is not an all-or-none endeavor. The disproportionate number of first-guesses and errors involving buttons directly adjacent to the correct button (i.e., approximate-trials) suggests that participants could often recall the approximate location of buttons when they couldn't recall the precise location. Thus, it may be the case that locations are learned in increasing levels of precision, such that as learning proceeds the precision increases.

Finally, because learning locations was the only alternative the no-label group had for improving performance, the performance curve for this group represents a pure curve for location learning in this paradigm (see Figure 4). Asymptote was not reached for this group until after 8 blocks (144 trials), indicating that acquiring the ability to reliably retrieve locations, like other chunks of information (e.g., a phone number), requires effort and practice. The nature of improvement in trial times for this group was well fit by a power curve ($r = .96$, see Figure 5d), further suggesting that location learning relies on the same learning mechanisms underlying other types of learning.

Search Cost

The results provide ample support for the Search Cost Hypothesis, which asserted that locations would be learned faster and more reliably in interfaces requiring a controlled search of the screen than in interfaces without such a requirement. The relevant group comparisons for evaluating this hypothesis are between the color-match group and all other groups. Support comes from the time disruption analysis, accuracy disruption analysis, and the location knowledge score analysis, all of which demonstrated that this group had weaker location knowledge than any of the other conditions.

In both the time and accuracy disruption analyses, the color-match group showed significant decreases in performance, made significantly more errors and took significantly longer to complete trials than the other groups. In the time analysis, the color-match-late group, unlike the other late groups, still showed significant disruption, indicating that participants in this group did not have accurate knowledge of all button locations even after 450 trials.

As compared to the meaningful and arbitrary groups, the color match group consistently performed more poorly on measures of location knowledge. The level of time and accuracy performance disruption was shown to be significantly higher than any of the other groups. With early and late location knowledge scores combined, the color-match groups had a score of 67%, significantly worse than either of the other label conditions (arbitrary = 94% and meaningful = 80%).

Evaluation Cost

The results also support the Evaluation Cost Hypothesis, which stated that users would rely more upon location in an interface with a high evaluation cost than in an interface with a low evaluation cost. Note that reliance upon location information requires that locations are not only known, but known to a degree that location acts as a sufficient cue to determine whether a given button is the correct one.

As indicated by Table 1, the anticipated search cost for the meaningful, arbitrary and no-label groups is constant, but the evaluation costs differ. As such, evaluating this hypothesis requires comparisons of the performance of these three groups in terms of the extent to which they learned and relied on location knowledge.

In terms of disruption, the meaningful group showed significantly more time disruption in performance than the arbitrary group, although there was no difference in accuracy disruption. The analysis of the location knowledge score (LKS), which included both verifies (trials in which participants accessed a ToolTip on the correct button prior to clicking it) and directs (trials where participants clicked the correct button without accessing a tip), revealed no differences between the meaningful, arbitrary and no-label groups. A more detailed analysis, which separated verifies from directs, revealed a significant difference between the meaningful and arbitrary groups in the number of verify trials. This analysis revealed that the meaningful text group was almost 3 times more likely to access a ToolTip on the correct button prior to clicking it, suggesting that this group was relying heavily on the label prior to the assessment block and less so on location knowledge. The arbitrary label group, which apparently had acquired the ability

to evaluate buttons based on location, was able to successfully evaluate whether the button was correct without resorting to a ToolTip.

Label vs. Location Learning

Although there is ample evidence that participants in the arbitrary condition learned and relied on button locations, the icon memory test showed that these groups also knew most of the icon-color associations by the end of the experiment. Clearly, the interaction between knowledge and skill in using an interface is just not a matter of learning and relying on location only or label only.

The learning curve analysis revealed a distinct advantage in the rate of location learning between the arbitrary and no-label groups, such that the arbitrary groups improved performance much more quickly than the no-label group. The most likely source of this advantage is that the icons, although not inherently meaningful, facilitated performance by enabling participants to distinguish the buttons from each other. Thus, participants in the arbitrary icon group, like the meaningful text and color-match groups, could eventually rely on the labels to evaluate whether or not the currently attended button was the correct one. The no-label group would have to access a ToolTip to perform such an evaluation.

Thus, it seems clear that the arbitrary group was relying both on label and location. What is not clear, however, is what was learned or relied on first. It is plausible that participants focused on the locations first and learned the labels as they completed trials (it would be difficult to ignore the labels completely). Alternatively, participants may have focused on the icon-color association first and then conducted a search similar

to one they would have conducted in the meaningful text condition (evaluate button labels until they find the one they are looking for, e.g., the book icon). The data collected in this experiment cannot resolve this issue.

The Need for Experiment II

Although Experiment I provided strong evidence in support of the hypotheses, there were several reasons for undertaking a second experiment. First, although careful control over experimental conditions allowed inferences to be drawn based on performance measures, the search cost hypothesis could be evaluated more directly using data derived from finer grained behavioral measures, such as point of gaze. Data on where participants were looking over the course of using the interface could provide additional evidence on whether or not participants are learning locations as well as insight on the process and rate of this learning (e.g., if participants make direct eye movements to the correct button instead of attending to multiple buttons, this is clear evidence that the location of that object has been learned).

Second, interesting questions emerged from Experiment I that could best be addressed with point of gaze data. The finding that there were a disproportionate number of approximate-trials suggested that approximate locations may be learned before the precise location. These data, however were based on accesses to ToolTips and errors and thus incorporated too few trials to enable an appropriate analysis. Further, these trials were taken only from assessment blocks, and thus may not have been reflective of more typical use. With point of gaze data, more accurate and

representative counts of approximate-trials over all trials and groups could be obtained in order to further investigate this phenomenon.

Third, data from the eye tracker could also be used to precisely extract search cost and evaluation cost from trial times, thus enabling quantifiable evaluation of the extent to which the four label conditions differ in the intended manner (i.e., did the intended pattern of search and evaluation costs emerge?). Recall that search cost is operationalized as the number of buttons that must be searched before the correct one is found, and evaluation cost is the amount of time required to determine if the currently attended button is the one currently needed. Both of these measures could be calculated from point-of-gaze data.

Finally, measures derived from eye-tracking data could be used to constrain the computational cognitive model built in the second, analytical, phase of this research. Instead of just fitting the model to trial time data, the model could be constrained to correspond to measures more indicative of location learning, such as number of buttons evaluated per trial.

METHODS - EXPERIMENT II

Experiment I provided strong evidence in support of the hypotheses, but the experimental method confined examination of location knowledge to specific snapshots of users' experience, i.e., the assessment blocks and the end of the experiment. To acquire more continuous, finer-grained measures of location learning, point of gaze data were collected in Experiment II.

Participants

Sixteen George Mason University undergraduates participated in the study for course credit. There was an equal number of participants in each of the four groups.

Materials

The layout of the experimental room was identical to Experiment I except for the inclusion of an Applied Science Laboratories model 504 eye tracker. This eye tracking system used a pan/tilt optics system in which the camera responsible for capturing the eye image sits facing the participant, as opposed to being mounted on headgear. Pupil diameter and point of gaze data accurate to 1 degree of visual angle were collected by the system and written to disk every 16 milliseconds.

The main experimental task was identical to the one used in Experiment I. At a viewing distance of 24 inches, each button subtended 1.4 degrees of visual angle and was separated from adjacent buttons by 3 degrees. The distance from the center rectangle to each of the buttons was 7.25 degrees. Given the 1 degree of accuracy of

the eye tracking system, this was designed to enable accurate identification of the current object of attention.

Procedure

The experimental procedure was a scaled-down version of Experiment I, excluding all but the color training, practice, main experimental task, and location memory test (see Appendix C for a schematic representation of the experimental procedure and full text of the protocol). Participants were led through a calibration procedure for the eye tracker, color training task, and practice tasks; they then performed the main experimental task for 16 blocks of 12 trials each, for a total of 192 trials (35% of the number of trials completed in Experiment I). The experiment contained only one between-subjects factor, label, which had the same four levels as in Experiment I (color-match, meaningful, arbitrary and no-label).

Dependent Measures

Although the performance measures used in Experiment I were also collected for this study, the primary measures of interest were derived from the eye-tracking data: the distance between the first-attended button and the correct button, the average button evaluation time per button per trial (corresponding to evaluation cost), and the number of buttons evaluated per trial (corresponding to search cost). The number of buttons evaluated per trial is also the primary measure of location learning in that it is assumed that as button locations are learned, the number of buttons evaluated per trial decreases. An additional measure derived from the eye data was the amount of time

spent attending to the colored rectangle at the beginning of a trial. The assumption underlying this measure is that participants are using this time to retrieve information about the currently needed button from memory (i.e., label, location). To the extent that participants were spending more time attempting retrieval, this is taken as a measure of the effort expended to avoid an extensive visual search of the screen.

The calculation of these measures required several key assumptions and a fair amount of post-processing. The first step involved segmenting and removing bad data from the eye-tracker output. The data were segmented into trials, where a trial is bounded by the point at which the colored-rectangle was initially presented and when the user clicked the correct button (error trials were excluded). Within a given trial, a data point was excluded if its reported pupil diameter was 0 (indicating a loss of data, typically a blink), or if either its reported x or y coordinate were 0 (again, indicating bad data).

The x and y coordinate units were then translated into pixels and plotted. With data from all trials for a given participant plotted simultaneously, it was possible to visually identify clusters of points for the colored rectangle and each of the 12 buttons (as well as their associated ToolTip rectangles). Cluster centroids were identified by centering the clusters over the mean of the distribution of points with the smallest standard deviation (see Figure 9). A zone was then constructed around each of these centroids. For the buttons, the zones were squares with sides of a length equivalent to 2.5 degrees of visual angle and for the center rectangle, the zone was a rectangle with a width of 5 degrees and a height of 2.5 degrees. The zone sizes were chosen based upon several constraints: (1) incorporating the 1 degree of error of the eye-tracking system,

and (2) minimizing the overlap between zones while maximizing the number of points assigned within zones.

A critical assumption in calculating the number of buttons evaluated per trial is the lower bound on the time required to evaluate whether or not a button is the one currently needed. A minimum in-zone time of 200 milliseconds was selected. This minimum is set primarily to filter out extraneous points passing through a zone (such as sometimes occurred after a blink) but to otherwise be inclusive. Once the zones and minimum criteria were defined, the number of buttons evaluated per trial was calculated as the sum of the visits to button zones. Multiple visits to the same button were counted only if the participant left the button's zone for longer than one second.

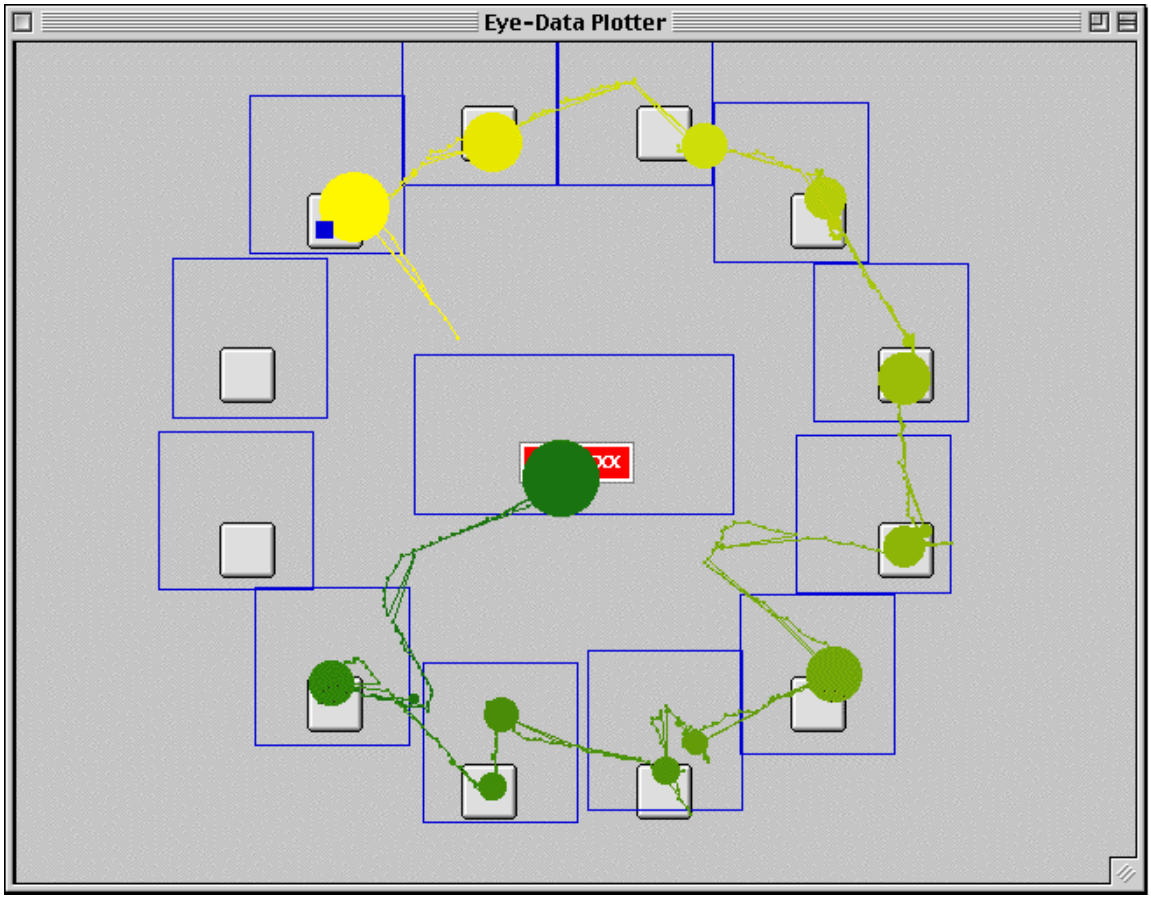


Figure 9. Example plot of eye data and analysis zones (rectangles surrounding buttons and rectangle in center) from a participant in the no-label condition. The filled circles represent sustained point of gaze such that the longer the gaze the larger the circle. The line represents the scan path and gets lighter as time passes through the trial.

Average evaluation time was calculated as the total time taken to complete the trial minus time spent in the center rectangle zone divided by the total number of button visits during the trial. The reason that this measure was used instead of simply summing the time spent in buttons zones and dividing by button visits stems from an inadvertent eye-tracking system setting that resulted in the reported eye position (i.e., x and y coordinates) being averaged over 500 milliseconds. Thus, a given x-y coordinate reported by the eye-tracker was a running average of the x-y coordinates over the

previous 500 ms. Because this averaging occurred across all conditions, the relative differences between groups, which constitute the basis for evaluating the hypotheses, are believed to remain unaffected. This issue is discussed further in Appendix D.

RESULTS - EXPERIMENT II

Performance Curves

As would be expected, a comparison between Figure 10, which plots trial time over blocks by label condition for Experiment II, and Figure 4, which plots the equivalent values for Experiment I, indicates a similar pattern of results. A 4 (Label) x 16 (Blocks) repeated measures ANOVA run on trial time yielded significant main effects of label, $F(3,12) = 9.11$, $p < .05$ and blocks, $F(15, 180) = 55.25$, $p < .05$, as well as a significant Label x Blocks interaction, $F(45, 180) = 12.91$, $p < .05$. The interaction was investigated in more depth via an analysis of simple main effects.

In terms of the level of performance improvement over blocks, the analysis revealed that whereas the arbitrary and no-label groups showed significant improvement over blocks, $Fs(15, 180) = 40.87$ and 51.38, respectively, $ps < .05$, the color-match and meaningful groups did not, $Fs(15, 180) < 1$ and $= 1.60$, respectively, ns. This pattern of results differs slightly from the Experiment I analysis in that the meaningful groups showed significant improvement in Experiment I but not in Experiment II. The difference is largely due to the number of blocks over which the analysis was run; this analysis includes 16 blocks of trials (192 trials) whereas the Experiment I analysis was only over 10 blocks (180 trials), such that stable performance over later blocks washed out the improvement in the earlier blocks.

The simple main effects analysis revealed no statistically reliable differences in trial time between the four groups after block 10, $F (3, 192) < 1$, ns. The pattern of

results of planned between-group contrasts run within the Label x Block interaction using these first 10 blocks is identical to that found in Experiment I. The no-label group took longer to improve than the arbitrary group, $F(9, 108) = 2.03$, $p < .05$, and the arbitrary group took longer than the meaningful text group $F(9, 108) = 17.03$, $p < .05$. The contrast between the meaningful text group and the color-match group was not significant, $F(9, 108) < 1$, ns, indicating that these two groups improved at roughly the same rate.

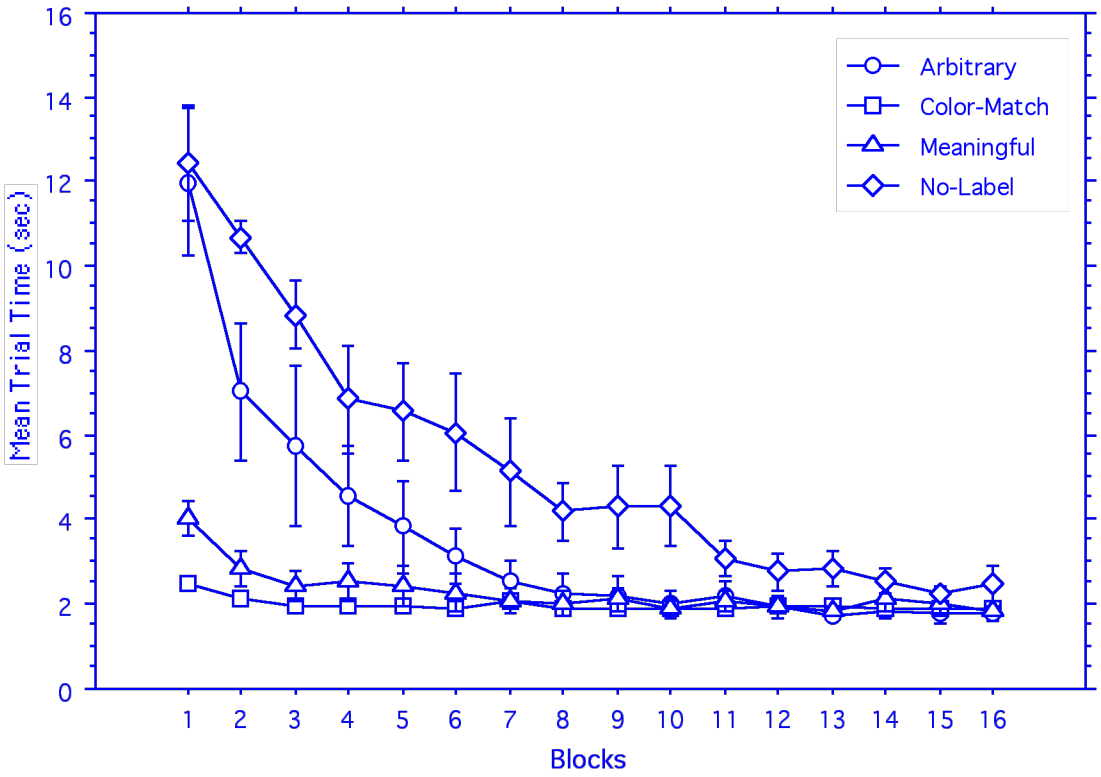


Figure 10. Trial times (in seconds) over blocks by label condition for Experiment II. Error bars indicate standard error.

Although the absence of the frequency manipulation in Experiment II obscures direct comparison with the times from Experiment I, the pattern of results is very similar, indicating that the effects are robust even with a significantly reduced sample size.

Average Button Evaluation Time

As indicated in Table 1, the label conditions were designed to vary evaluation cost, i.e., the time and effort required to determine if the currently attended button is the one currently needed. To evaluate the effectiveness of this manipulation, a 4 (Label) x 16 (Blocks) repeated measures ANOVA was run on average evaluation time. This analysis yielded significant main effects of label, $F(3,12) = 10.61$, $p < .05$, and Blocks, $F(15, 180) = 22.90$, $p < .05$, both of which were superceded by a significant Label x Blocks interaction, $F(45, 180) = 9.35$, $p < .05$, shown in Figure 11.

As can be seen in this figure, the color-match and meaningful groups show similar, stable performance over blocks. A comparison of the means for these groups over blocks reveals no difference $F(15, 180) < 1$, ns. Thus, the time to evaluate the text label is equivalent to the time needed to match the rectangle color with the labels in the color-match condition. Indeed, the meaningful text group appears to be faster at block one, though this difference is not significant, $M_{meaningful} = 0.61$, $M_{color\text{-}match} = 0.97$, $t(1)= 2.65$, $p < .05$. The value for the color-match group is most likely larger than the meaningful condition due to mouse movement time, which is implicitly included in calculation of the average button evaluation time. In trials where only one or two buttons are attended, such as in the case of the color-match group and the other groups in later blocks, the time to get the mouse to the button becomes the lower

bound for average evaluation time. Data from the mouse movement test provides a time estimate of 720 milliseconds. The mean times over all blocks for the color-match and meaningful groups are very close to this bound, $M_{color\text{-}match}$ = 770 ms and $M_{meaningful}$ = 728 ms.

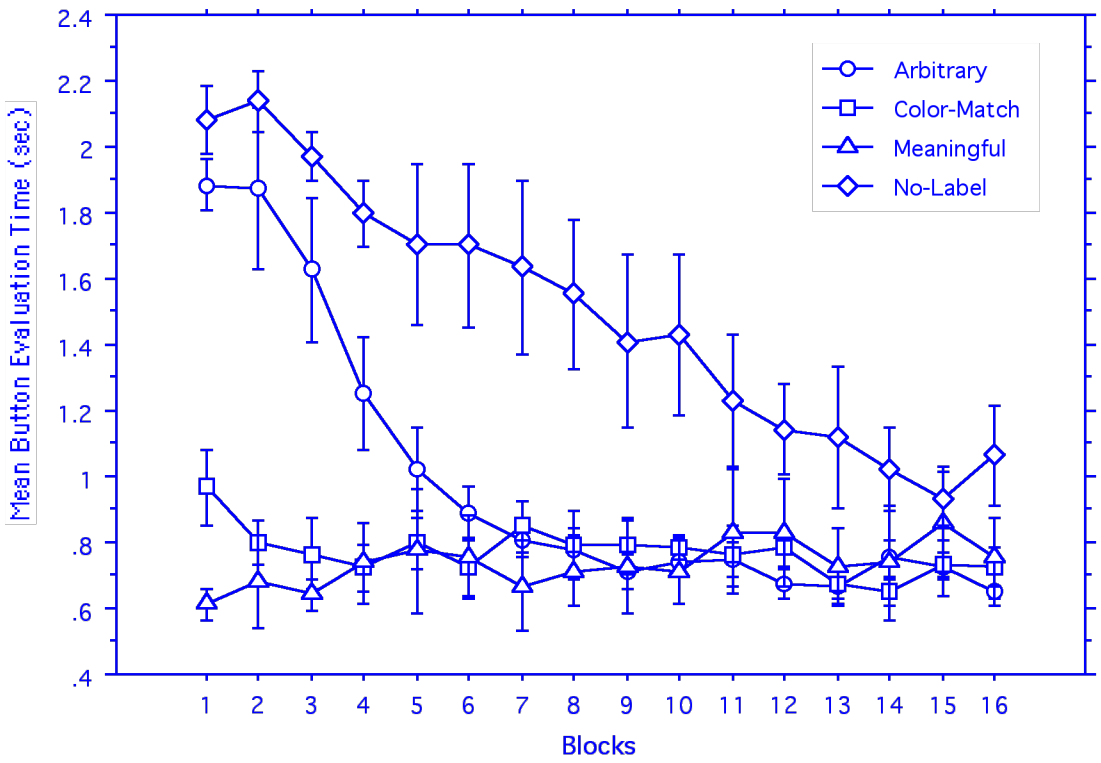


Figure 11. Average evaluation time per button (in seconds) by block and label. Error bars indicate standard error.

The no-label group spent more time evaluating buttons than the arbitrary group, $F(15, 180) = 3.49$, $p < .05$, and the arbitrary group, in turn, spent more time than the meaningful group, $F(15, 180) = 16.78$, $p < .05$. As can be seen in Figure 11, the no-label group shows slow, gradual improvement as compared to the arbitrary group. The

arbitrary group reaches the same level as the meaningful and color-match groups by block 7. The difference between the arbitrary and no-label groups, attributable to the no-label group's sustained reliance on ToolTips (see Figure 12), again demonstrates the usefulness of even meaningless labels. Even this weak visual cue can serve to enhance participant's recall of the color associated with the button, and thus enable evaluation of the button without necessitating a ToolTip.

Rectangle Study Time

The amount of time spent looking at the colored rectangle at the beginning of a trial was also calculated and analyzed. As can be seen in Figure 13, which plots rectangle study time by group over blocks, the differences between groups in time spent looking at the center are qualitatively similar to the differences the trial times plotted in Figure 10. A 4 (Label) x 16 (Blocks) repeated measures ANOVA yielded a non-significant main effect of Label, $F(3,12) = 2.50$, ns, a significant main effect of Blocks, $F(15, 180) = 13.19$, $p < .05$, and a significant Label x Blocks interaction, $F(45, 180) = 2.71$, $p < .05$. This interaction was investigated in more depth via contrasts and simple main effects.

The analysis of simple main effects revealed that whereas the arbitrary, meaningful, and no-label groups showed significant improvement over blocks, Fs(15, 180) = 7.49, 2.17 and 11.5, respectively, ps < .05, the color-match group did not, $F(15, 180) < 1$, ns. Thus, the three controlled search groups spent increasingly less time over blocks attending to the center rectangle, whereas the color-match group spent a relatively constant amount of time. To the extent that the time spent looking at the rectangle was being used by the participants to attempt retrieval of the label and

location of the currently needed button from memory (in lieu of conducting a visual search), the decrease in rectangle study time by the controlled search groups may reflect learning of the labels and/or locations. The underlying assumption here is that as participants acquired the ability to quickly retrieve the required information (i.e., the label or location of the currently needed button) they spent less time looking at the rectangle. The color-match group, to the extent that they were relying on the pre-attentive search to locate the correct button, did not have to bother with such retrievals.

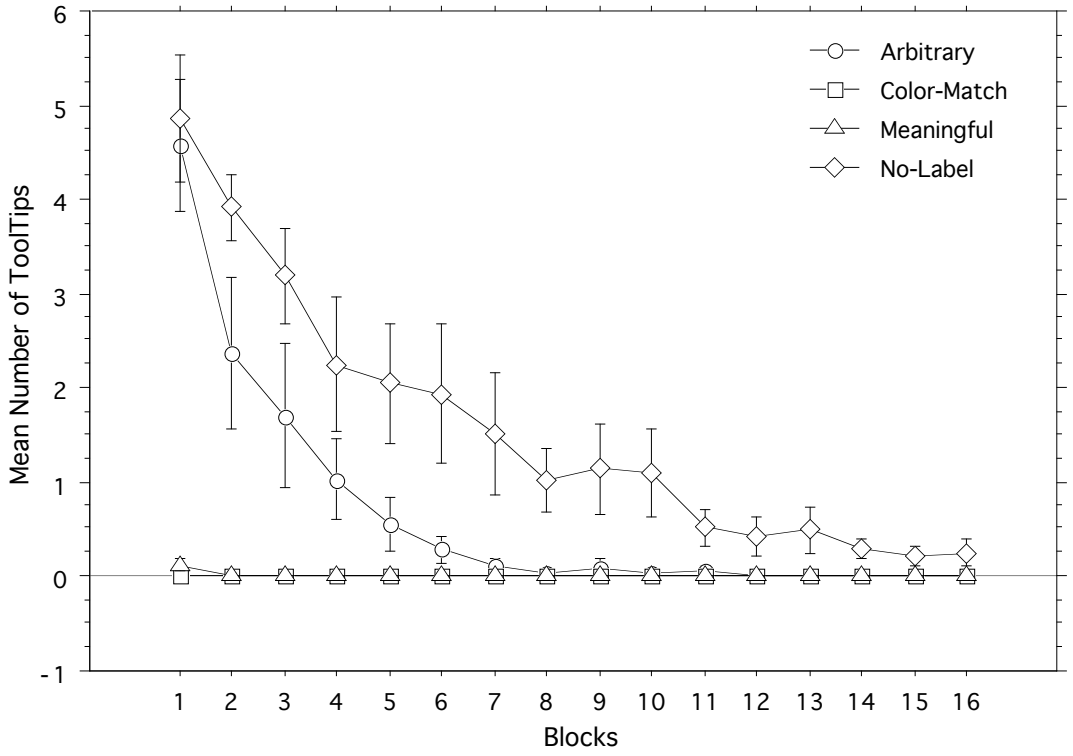


Figure 12. Mean number of ToolTips accessed per trial by block and label. Error bars depict standard error.

Contrasts between the groups within the Label x Block interaction revealed some differences in the rate of the decrease in study time. The meaningful group decreased study time faster than the no-label group, $F(15, 180) = 2.90$, $p < .05$, but not faster than the arbitrary group, $F(15, 180) = 1.64$, ns. There was no difference between the arbitrary and no-label groups, $F(15, 180) < 1$, ns, indicating that these two groups improved at roughly the same rate.

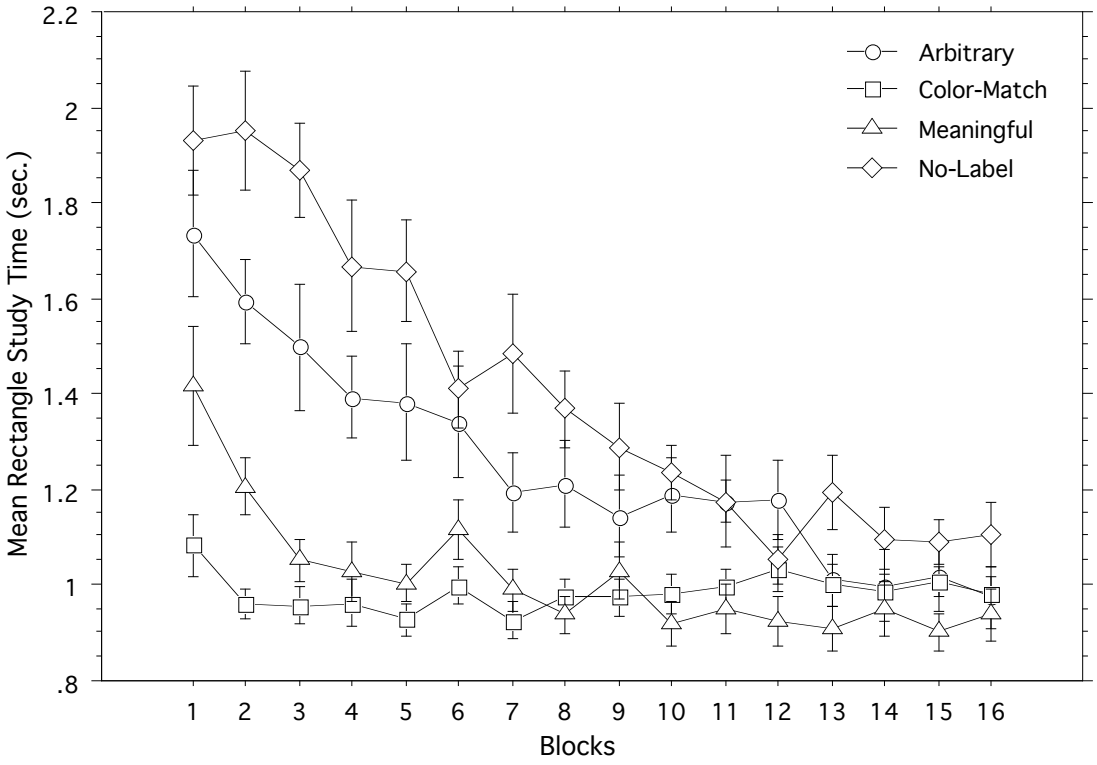


Figure 13. Mean time spent looking at the colored rectangle per trial by block and label. Error bars depict standard error.

Number of Buttons Evaluated

Figure 14 shows the number of buttons evaluated per trial by block and condition. There are two primary features apparent in this graph. First, the arbitrary, meaningful and no-label groups appear to start off at the same point in block 1 ($M_{arbitrary}$ = 5.27, $M_{meaningful}$ = 4.94, $M_{no\text{-}label}$ = 5.06), whereas the color-match group starts much lower, $M_{color\text{-}match}$ = 1.57. The second feature is that the arbitrary, meaningful, and no-label groups show a strong and similar curvilinear trend while the color-match group remains quite flat over blocks. The first of these features, the starting points of the groups, underlies the distinction between the pre-attentive search (low search cost) enabled by the color-match condition and controlled search (high search cost) required by the other three conditions. To determine the statistical reliability of this distinction, a one-way ANOVA was run with label as the independent variable and number of buttons evaluated for block 1 as the dependent variable. This analysis yielded a significant main effect of label, $F(3, 12) = 7.60$, $p < .05$, and thus allowed contrasts to be run. These contrasts revealed a significant difference between the pre-attentive (color-match) group and the three controlled search groups, $M_{color\text{-}match}$ = 1.57, $M_{controlled\text{-}search}$ = 5.09, $F(1, 12) = 22.43$, $p < .05$, but no significant differences between any of the controlled search groups: arbitrary versus meaningful, arbitrary versus no-label or meaningful versus no-label, $Fs(1,12) < 1$, ns. This analysis supports the validity of the distinction and thus, the effectiveness of this manipulation.

The second feature, the nature of the decrease in number of buttons evaluated over blocks, is taken as the primary measure of location learning. A 4 (Label) x 16

(Blocks) repeated measures ANOVA run on buttons-attended yielded a non-significant main effect of label, $F(3,12) = 2.97$, ns, a significant main effect of Blocks, $F(15, 180) = 38.17$, $p < .05$, and a significant Label x Blocks interaction, $F(45, 180) = 3.85$, $p < .05$. This interaction was investigated in more depth via contrasts and simple main effects.

The simple main effects revealed that only the color-match group failed to show a significant decrease in the number of buttons evaluated over blocks, $F(15, 180) < 1$, ns. Planned contrasts conducted on the Label x Blocks interaction revealed no significant differences between any of the controlled search groups, Fs$(15, 180) < 1$, ns. Thus, as measured by acquisition of the ability to accurately move attention to the correct button, the three controlled search groups learned locations at the same rate. The color-match group, unlike the other groups, demonstrated the ability to find and move attention to the correct button from the very first block.

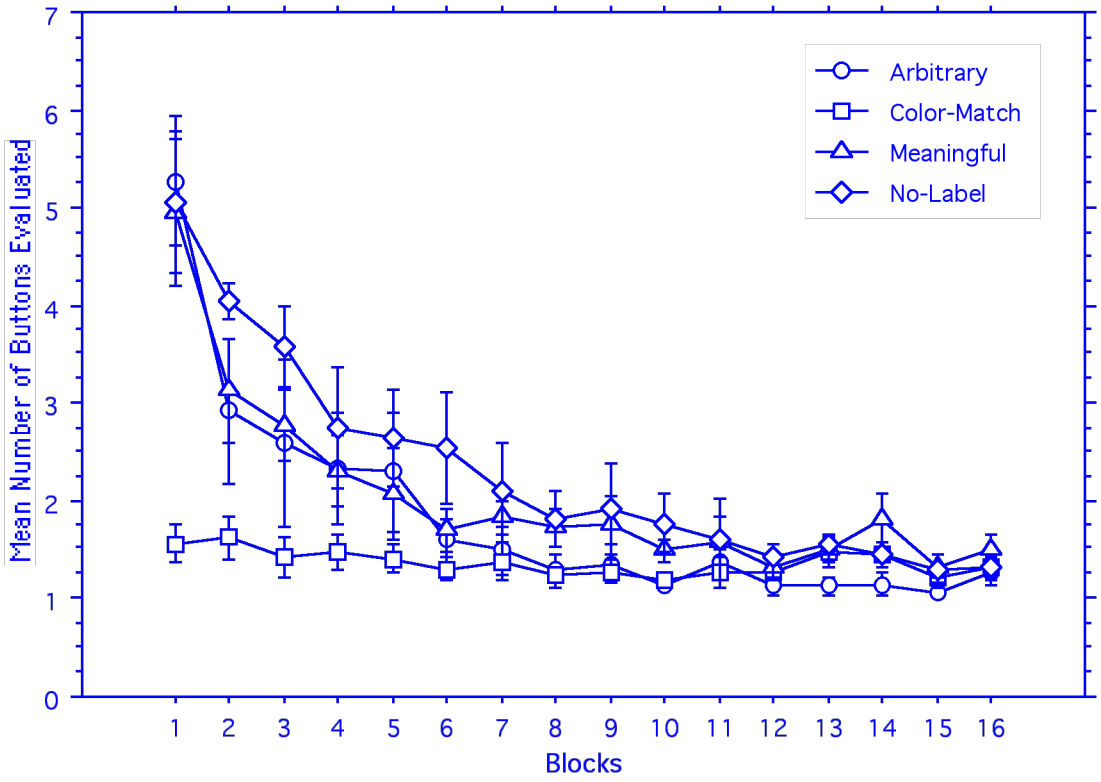


Figure 14. Mean number of buttons evaluated by block and label. Error bars depict standard error.

A look at Figure 14 reveals that the lowest average number of buttons evaluated for any group is about 1.5. This value would be expected to be closer to one for all blocks in the color-match group and for later blocks in the other groups, once locations had been learned. This unanticipated result is explored further in the next section.

Proximity Analysis

In Experiment I, participants were shown to have a disproportionately high number of errors and tip-accesses on buttons directly adjacent to the correct button, called approximate-trials. The collection of eye-position data for Experiment II enabled

the exploration of this phenomenon at a much finer grain. For each trial, the distance from the first-attended button to correct button was calculated in the same manner as it was for Experiment I (see Figure 8). Approximate-trials were then identified as those in which: (1) the first-attended button was adjacent to the correct button and, (2) the correct button was the only other button attended and clicked. Figure 15 plots the percentage of approximate-trials beside values for other trials where the first-attended button was not the correct button. As can be seen in this figure, the frequency of approximate-trials, at 35.7%, is roughly three times higher than for other trials where participants did not look at the correct button first (i.e., button distance is not zero).

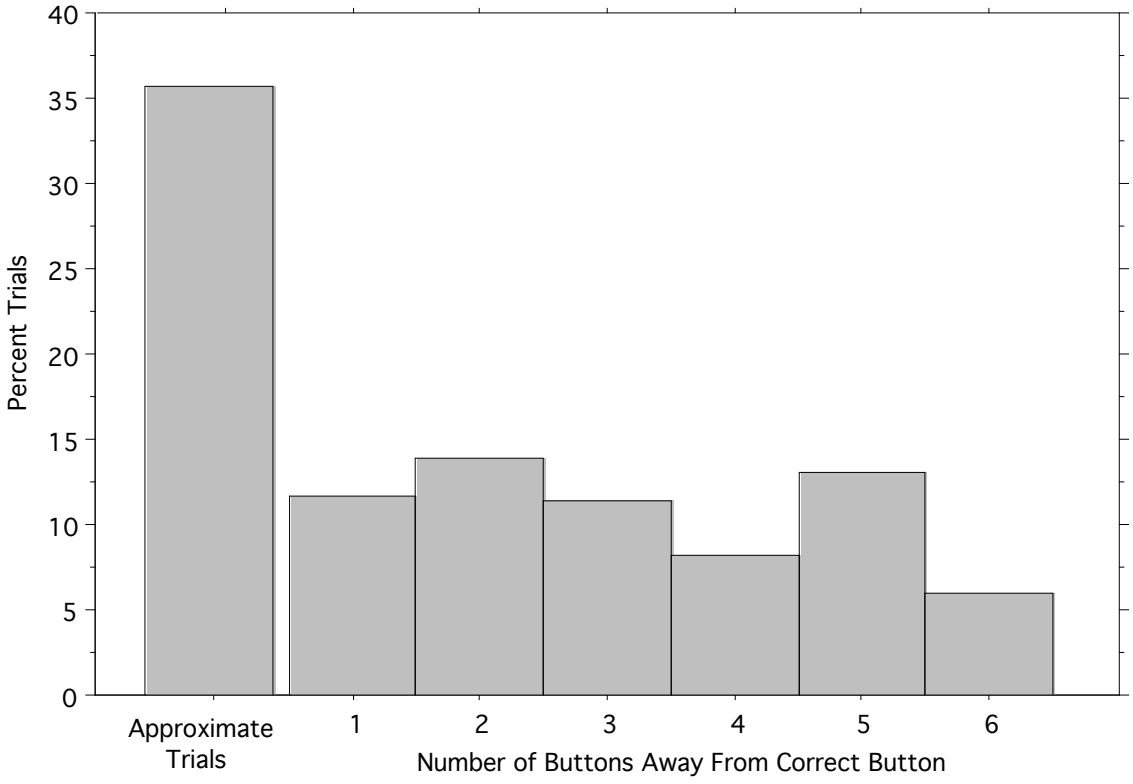


Figure 15. For trials in which the first-attended button was not the correct button, the distribution of distances (in number of buttons, see Figure 8) from the first-attended button to the correct button. Approximate-trials are distinct from the reported 1s in that to be counted as an

approximate-trial, the participant had to attend to only two buttons: the adjacent button, then the correct button. Thus, 1s include trials in which more than two buttons were attended; the implicit assumption being that the participant attended to an adjacent button merely due to chance.

Although the percent of approximate-trials is lower than the 55.5% found in Experiment I, it is important to point out that the circumstances under which button distance was assessed is quite different. In Experiment I, these were only measured in the assessment blocks, so there were a maximum of 12 possible measurements per participant as compared to Experiment II, where button distance was assessed for each of the 192 trials. With measurements for each trial, it is possible to determine if there is a pattern of approximate-trials over blocks.

Figure 16 plots the proportion of approximate-trials by block and label condition. The overall mean proportion of approximate-trials is 12%, or about 1.5 trials per block. It is difficult to discern much of a pattern from this figure, although it appears that the meaningful and no-label conditions show an initial increase in the proportion of approximate-trials, the color-match and arbitrary group appear to show a decrease. A 4 (Label) x 16 (Blocks) repeated measures ANOVA yielded only a significant main effect of label, $F(3,12) = 4.45$, $p < .05$. Pairwise contrasts on the main effect revealed significant differences between the arbitrary ($M = .09$, $SD = .10$) and both the meaningful ($M = .15$, $SD = .10$) and the no-label groups ($M = .13$, $SD = .08$), and also between the color-match group ($M = .10$, $SD = .08$) and the meaningful group, Fs(15, 180) = 11.11, 5.69, and 5.97, respectively, ps $< .05$.

If there is something systematic about the relationship between the proportion of approximate-trials over blocks and location learning, it is not readily apparent from these

data. If participants are first learning the approximate location of the buttons and then subsequently honing their accuracy, the tail ends of the curves for the controlled search groups should become increasingly stable and close to zero. These curves, however, are noisy but overall relatively flat, and hovering at an average of about one approximate-trial per block, suggesting that the approximate effect may instead stem from noise in eye movements or imprecise location memory.

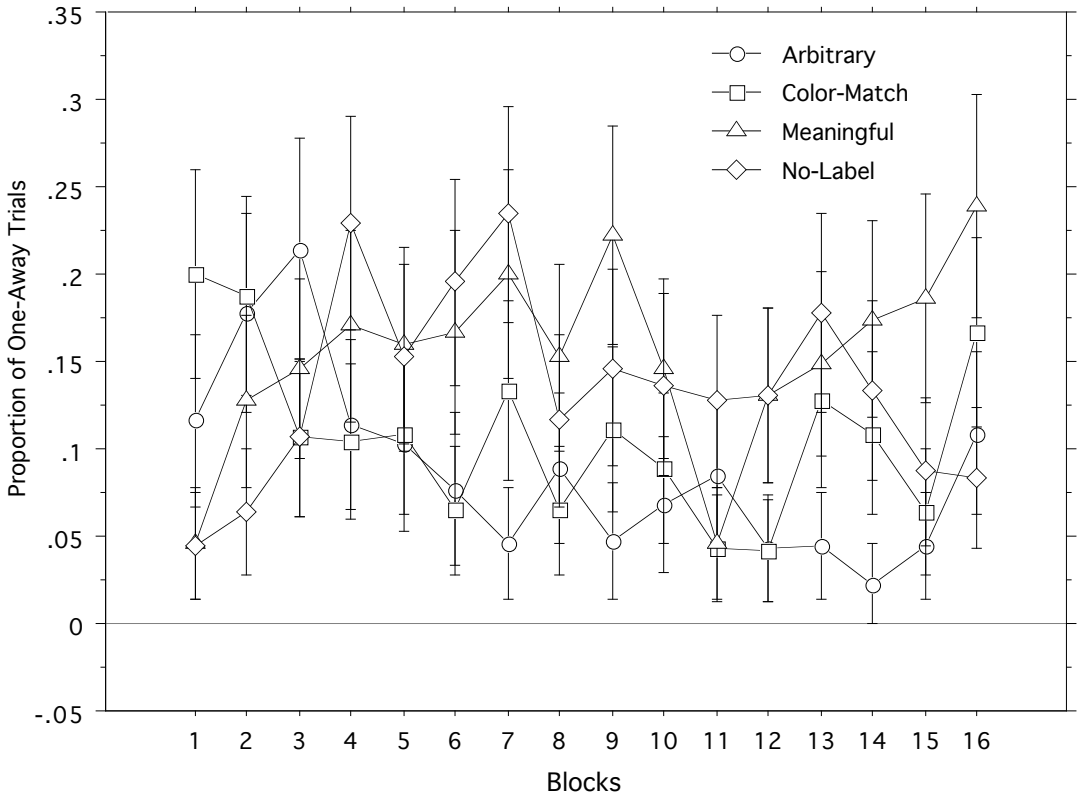


Figure 16. Proportion of approximate-trials versus all trials by block and condition. Error bars represent standard error

Support for the noisy eye-movement explanation comes from the lack of difference between the color-match group and the combination of the controlled search groups ($M_{color\text{-}match}$ = .10, $M_{controlled\text{-}search}$ = .12), F (1, 12) = 1.55, ns. To the extent that the color-match group is using a pre-attentive search (i.e., not relying on memory), the approximate-trials for this group should be attributable to error in eye movement error rather than noise in location memory. To the extent that participants in this experiment are emphasizing speed over accuracy in making the saccade from the rectangle to the desired button, there is some support for this explanation in the eye movement literature. Several studies have demonstrated that saccade accuracy significantly increases when additional time is provided for planning the movement (Coeeffe & O'Regan, 1987; Viviani & Swensson, 1982). This speed/accuracy trade-off explanation relies on the assumption that the participants have accurate knowledge of where the saccade is to be directed. In the color-match group, the accuracy of the knowledge would be based on the automatic detection of the button. Because in the other label conditions, this would require retrieving the location from memory, the data should initially show chance levels of approximate-trials followed by an increase and eventually steadying over later blocks. Although, the Label x Block interaction was not significant, the controlled search groups do appear to show an increasing trend over the first 5 blocks of trials.

The data from Experiment I seem counter this explanation, however. Recall that the approximate counts from Experiment I were based on error trials and trials in which participants accessed ToolTips. If the effect was merely the result of error in eye movement, participants would likely not have clicked or waited for a tip. In addition, an

explanation based on the premise that the location knowledge is noisy or approximate could also explain these features in the data, at least for the controlled search groups. Thus, though it is not clear form this data, it is plausible that the approximate effect is the combined result of noise in location memory and eye movement.

Location Memory Test

Location memory was scored in terms of mean absolute distance in pixels from the correct location, as it was in Experiment I. Means and standard deviations are reported in Table 8. The overall mean error for Experiment II, at 54.02 pixels, is larger than the 29.59 pixels reported in Experiment I. This poorer performance is to be expected, given that participants in Experiment II: (1) were not forced to rely on location knowledge by having the button labels unexpectedly removed and (2) completed about one-third of the number of trials as their counterparts in Experiment I.

Table 8. Means, and standard deviations for accuracy scores on 1ocation memory test. The scores are the average euclidean distance in pixels from the guessed locations to their correct respective button location (n = 4).

Condition	M	SD
Color-Match	99.29	29.49
Meaningful	45.86	3.96
Arbitrary	29.43	8.66
No-Label	41.51	13.39
Overall	54.02	

A one-way ANOVA run on error score with group as the between-subjects factor yielded a significant effect of label, $F(3, 12) = 13.47$, $p < .05$. As can be seen in Table 8, the error for color-match group, at 99.29 pixels, is more than twice that of any other group, indicating that this group had less accurate location knowledge by the end of the experiment. Pairwise contrasts confirm that the score for the color match group was significantly worse than that of the arbitrary, meaningful and no-label groups, $Fs(1, 12) = 34.28$, 20.05 and 23.45, respectively. No other pairwise contrasts were significant, indicating that the arbitrary, meaningful and no-label groups had equally accurate location knowledge by the end of the experiment.

DISCUSSION - EXPERIMENT II

The results of Experiment II verified that the search and evaluation cost manipulations had the intended effects. The effectiveness of the search cost manipulation (see Table 1), was demonstrated in a block one comparison between the controlled search groups, who on average attended to about five buttons per trial, and the color-match group, who only attended to about one button. Given the operational definition of search cost as the number of buttons which would have to be attended prior to locating the correct one, the search cost was indeed higher for the meaningful, arbitrary, and no-label groups than for the color-match group.

In terms of the hypothesized effects of search cost on location knowledge (i.e., the search cost hypothesis), the results of the location memory test revealed that the high search cost groups had more accurate location knowledge than the low search cost group. Thus, in congruence with this hypothesis, the increased search cost imposed by the interface led to faster acquisition of location knowledge.

The evaluation cost manipulation (see Table 1) was largely verified via the analysis of average button evaluation time. As intended by the manipulation, participants were shown to have spent more time evaluating each button in the no-label condition (high cost) than in the arbitrary (moderate cost) condition, and likewise more time on the arbitrary than the meaningful (low cost) condition. There was no difference between the color-match and the meaningful conditions, although the lack of a difference is likely attributable to the mouse movement time washing out whatever differences may have

existed between these groups. Alternatively, it is possible that it took participants no longer to evaluate a text label than a blotch of color.

Finally, the approximate-trial phenomenon found in Experiment I and replicated in Experiment II was explored in more detail. Although the results failed to clearly identify the cause of this phenomenon, they did raise some interesting alternative explanations including the presence of a speed/accuracy trade-off in eye-movements and the retention of only approximate location knowledge for use in the search phase of the task. Future studies will have to try to tease these explanations apart.

CONCLUSIONS - EXPERIMENTS I & II

The results from Experiments I and II demonstrate that location learning is pervasive. The participants in all groups learned the button locations to some degree. This finding is congruent with findings in the HCI, problem solving, and reading literatures described above, all of which demonstrated that people learn locations with practice. The results of the present research extended these findings by demonstrating that location learning is not only pervasive, but also subject to the cost structure of the interface. Consistent with the search cost hypothesis, as the cost of using the label to locate the currently needed button (i.e., search cost) increased, so did the rate of location learning and reliance on location knowledge. Consistent with the evaluation cost hypothesis, as the cost of using the label to evaluate whether a given button is the one that is currently needed (i.e., evaluation cost) increased, so did the reliance on location knowledge in the evaluation phase of task performance. Thus, support was found for both hypotheses.

The results from both experiments indicate that users of direct manipulation interfaces will rely on location as a performance cue only to the extent that the interface provides them with no less-effortful alternative. Embedded in this hypothesis is the assumption that the user interacts with the interface in a rational manner, choosing the interaction strategy that incurs the lowest cost. Along with the data of Experiments I and II, this assumption was used to guide the development of a computational cognitive

model of learning and performance on the experimental task. The model, built using ACT-R/PM, is described in the section that follows.

ACT-R/PM MODEL

This section presents the second, analytical, phase of this two-phased research effort. Data from the first, empirical, phase were used to constrain the behavior of a computational cognitive model built using ACT-R/PM (Byrne & Anderson, 1998). The model interacts with the same interfaces as the participants. As such, its performance is constrained to coincide with key attributes of participants' behavior, from fine-grained components of interaction such as eye and mouse movements to higher order measures such as the decreased performance time resulting from location learning. I begin with an overview of the relevant components of ACT-R/PM, then describe the model and its fit to the data, and conclude with a discussion of the implications of the model for a more general account of location learning.

ACT-R/PM Overview

ACT-R (Anderson & Lebiere, 1998) is a production system-based cognitive architecture (Newell, 1990) that assumes two kinds of memory: declarative memory and procedural memory. Declarative memory contains chunks of factual knowledge and past goals. Each chunk is a member of some category and has some number of slots, which can contain other chunks. Procedural memory contains production rules of the form: IF <goal-specification> and <optional-memory-retrievals> THEN <goal-action>. The primary

control structure is a last-in, first-out goal stack with the top goal determining which production rules are candidates to be fired.

New declarative memory chunks are added to the system by virtue of either being popped off the goal stack or being encoded from the (simulated) external environment. Whenever a chunk is about to be created that is identical to a chunk currently in memory, these chunks are merged together. This merging process means that there will be only one canonical chunk in memory for a given piece of declarative knowledge.

A recent extension to ACT-R, the perceptual-motor component (Byrne & Anderson, 1998), allows for actions such as eye and mouse movements to occur in parallel with cognition. One primary benefit of this component is that it enables ACT-R models to interact with the same software as that used by human participants, thus enforcing constraints on the flow of information and execution of interface actions.

The primary ACT-R/PM sub-component implicated in location learning is the vision module, so the process of moving attention to screen objects will be discussed in some detail. This process begins with a production rule that sets off a search for an object with specific features (e.g., color, shape) in the visual icon. The visual icon is a structure containing low-level descriptions of all objects contained in the model's visual task environment (i.e., the computer display). This process, if successful, creates a declarative memory chunk representing the location of a candidate object. If a chunk representing that location already exists in declarative memory, the module merges that chunk instead of creating a new one. The resulting visual-location chunk can then be used as an argument in an instruction to move visual attention to that location. After a

delay of 185 milliseconds (which is meant to simulate the time required to move visual attention), a chunk representing the object itself is created and thus made available to the production rules in procedural memory.

Activation Learning

At any given time, each chunk in ACT-R's declarative memory has an activation level associated with it which is the sum of its base level activation (which reflects the log prior odds that a chunk is needed) and the activation being spread to it from the current activation sources (which are the chunks contained in slots of the current goal and slots of the current object of visual attention). The activation being spread from the activation sources is calculated based on the associative strength from each activation source to each chunk that the production is trying to retrieve from memory. This associative strength represents the log likelihood that the chunk being retrieved is needed given that the activation source is in the goal context. Activation levels are a critical determinant in the behavior of a model in that if a chunk's activation is below a threshold value, the chunk cannot be retrieved and the production rule fails.

Both base level activation and associative strength can be learned. In base level learning, the base level activations of chunks increase each time a chunk is successfully retrieved or merged, but decrease as a function of decay over time. This mechanism works such that base level activation for a given chunk increases as a power function of practice and decreases as a power function of delay between uses.

Associative strength learning is best described in the form of an example: let chunk AS be an activation source (e.g., in a goal slot) and chunk MR be a chunk to be retrieved (as specified in a production rule). The associative strength between AS and

MR depends on a number of factors, including: the number of chunks that contain AS, the number of times MR has been retrieved when AS was an activation source, and the number of times AS has been an activation source. The interassociative strength between AS and MR increases as a function of the number of times AS has acted as a retrieval cue for MR, but decreases relative to the number of other chunks which contain AS and the number of times AS has been an activation source. This decrease in associative strength relative to the number of chunks which contain AS predicts a fan effect. In essence, this means that an activation source which is uniquely associated with a chunk to be retrieved will act as a faster and more reliable cue for retrieval than a cue which is used often and is common to many chunks.

As a summary example, if chunk MR is used often it will have a high base-level activation and therefore a higher probability of being successfully retrieved. Further, if AS is often an activation source when chunk MR is retrieved and is uniquely associated with MR, then MR will have an even higher probability of being successfully retrieved when AS is an activation source. As the lag between retrievals of chunk MR increases or as the number of other chunks which contain AS increases, the probability of successful retrieval of MR decreases.

Search versus Retrieval

It is possible to have multiple production rules whose conditions match the current goal. In this case, the expected gain of each rule determines which rule will be tried. The expected gain is a function of the probability of successfully achieving the goal, the cost (in time), and the value of the goal. If the rule attempted first fails to retrieve the memory chunks specified in its condition side, then the rule itself fails and

the next one is tried. This arrangement enables a compute versus retrieve modeling paradigm which has been used to model the learning of arithmetic facts (Lebiere & Anderson, 1998). In this paradigm, there are two primary rules in direct competition with each other: one rule which simply retrieves the answer from memory (i.e., from a past problem-solving episode) and the other which sets out to compute the answer (e.g., counting on fingers). The retrieve rule is preferred because it has a higher expected gain, but is unable to fire initially due to the inability to retrieve the facts from memory (either due to non-existence of the fact or sub-threshold activation). As the facts are created via the compute production, they are merged into a single chunk that gets a base-level and associative activation boost, thus increasing the probability of retrieval in the future. Eventually, the activation spread reaches a threshold level, such that the retrieve production succeeds and the compute production no longer fires.

The model of location learning described below uses a variant of this paradigm both in the search phase and the evaluation phase of task performance. In the search phase, the retrieve rule, instead of retrieving an arithmetic fact chunk, attempts to retrieve a past goal in which the currently needed button was clicked, as well as the visual-location chunk associated with the button, and instead of computing an answer on retrieval failure, the competing rule sets out to conduct a visual search of the display to find the desired button. Likewise, in the evaluation phase, if the model cannot retrieve the knowledge required to determine if the currently attended button is the one currently needed, the model will instead rely on a ToolTip to determine the color associated with the button. The details of the model are described below.

Model Description and Fit to Data

There are 35 production rules to model the four conditions. Each condition requires a subset of the rules ranging in size from 11 rules for the color-match condition to 27 rules for the arbitrary condition. Twenty percent of the rules are common to all conditions and the overlap between conditions ranges from 23% for the color-match and arbitrary label conditions to 86% for the arbitrary and no-label conditions.

The model interacts directly with the experimental software used by the participants; thus, the model specifies movements of the cursor, mouse clicks, and movements of visual attention to objects on screen. For the data reported below, the model was run 10 times for each of the four conditions. Data on five aspects of the model's performance, including: performance time, number of buttons attended, average button evaluation time, number of ToolTips accessed, and rectangle study time, were collected and written to a log file. These above measures were used to assess the fit to the human data.

As described above, the model relies on a search versus retrieve paradigm, so the ability to retrieve a needed chunk is a critical component of the model. Because acquiring the ability to retrieve a needed chunk (i.e., learning) is accomplished via the adjustment of chunk activation values in ACT-R, both base-level activation and associative strength learning were enabled for the model runs. To review, base-level activation refers to the probability that a chunk is needed and is sensitive to usage patterns such that it increases with each use and decreases with delay between uses. Associative activation provides an adjustment to base-level activation to make it

sensitive to the current (goal or external) context, e.g., retrieving the chunk representing a mailbox icon is more likely when a visual representation of a mailbox is the focus of visual attention.

Table 9. ACT-R parameter values used in the model and their associated defaults (in parentheses).

Parameter	Value	Brief Description
Activation Noise [:ans]	0.7 (0.5)	Cycle to cycle noise added to activation values of chunks
Associative Learning [:al]	1.0 (1.0)	Weighting of prior associative strengths
Base-Level Learning [:bll]	0.3 (0.5)	Decay and learning rate for base-level activation
Expected Gain Noise [:egs]	0.1 (0.0)	Cycle-to-cycle noise added to the expected gain of an instantiation
Latency Factor [:lf]	2.0 (1.0)	Scaling factor mapping activation values to latency
Retrieval Threshold [:rt]	2.3 (0.0)	Minimum activation value for successful chunk retrieval

The ACT-R parameter settings, summarized in Table 9, were held constant for all conditions. In the interest of limiting degrees of freedom required for the model's fits, an attempt was made to leave parameter values at their defaults. When the defaults were inadequate, values were adjusted with the twin goals of minimally deviating from the default value and simultaneously fitting all five performance measures described above for all four conditions. Given that there were 16 blocks of trials, four conditions, and five performance measures being modeled, the parameters in Table 9 were used to fit 320

data points. For a more detailed description of the parameter values, the interested reader is pointed to Appendix E1.

The description of the model is organized by the major phases of task performance: the search phase and the evaluation phase. The discussion will refer to various tables that contain descriptive versions of the key production rules. These rules are presented in ACT-R/PM syntax, along with brief descriptions, in Appendix E. To provide a high-level overview, a summary of the key assumptions underlying the model's behavior is presented in Table 10.

Search Phase

Encode Color and Get Label

When a trial begins, the model first encodes the color of the rectangle and pushes a goal to locate the button that applies that color. Once the goal color has been encoded, the model attempts to put the chunk representing the label of the currently needed button in the locate goal, unless the model is running in the color-match condition. This is not attempted in the color-match condition because the goal color itself represents the label of the currently needed button (i.e., they are the same chunk), so a separate procedure for encoding the label is not performed.

In the meaningful condition, the model fires a rule named MN-RETRIEVE-LABEL that gets the name of the goal color from the goal color chunk and stores it in the locate goal. This rule always succeeds, under the assumption that, for the participants, the name of the color is closely associated with color itself (i.e., the word red is highly associated with the perceptual experience of red).

Table 10. Summary of the behavior of the model in the four conditions for the two phases of performance. Values in parentheses refer to the section of Appendix E containing the syntax.

Search Phase

ENCODE COLOR (E4)

- All conditions begin by encoding the color of the rectangle

GET LABEL (E5)

- The color-match condition skips this phase as the rectangle color and label are represented by the same chunk
- The meaningful condition always succeeds in accessing the name of the rectangle color
- The arbitrary condition attempts to retrieve the label of the button that applies the rectangle color – these retrievals fail at first but eventually succeed in later blocks
- The no-label condition always encodes the fact that the button is blank

DETERMINE LOCATION (E6)

- The color-match condition relies on the pop-out effect to locate the currently-needed button
- The meaningful, arbitrary and no-label conditions attempt to retrieve a past use of the currently needed button and its location before undertaking a search of the screen
- The process of moving attention to a button location is noisy for all conditions, such that on 12% of the trials, visual attention lands on a button adjacent to the intended one

Evaluation Phase

ENCODE LABEL (E7)

- All conditions first encode the label on the button currently being attended

COMPARE (E8)

- The color-match condition uses a label-matching strategy, comparing the label on the current button with the rectangle color
- The meaningful condition uses a label-matching strategy, comparing the label on the current button with the label previously retrieved
- The arbitrary condition, if the label was previously retrieved, uses a label-matching strategy, comparing the retrieved label with the current label
- The no-label condition and arbitrary condition (if label not retrieved) use a location-recognition strategy, attempting to retrieve a past use of the currently attended button and its location, and comparing the color it applied with the rectangle color - if cannot retrieve past use of button waits for ToolTip and compares ToolTip color to rectangle color

TRY AGAIN (E9)

- The color-match condition conducts another pre-attentive search for the correct button
- The meaningful condition, on 30% of the trials encodes a button chunk representing the current (wrong) button - the arbitrary and no-label conditions do this on 50% of the trials
- The controlled search conditions on half of the trials attempt to retrieve the location of the currently needed button and on the other half move attention to the nearest unattended button.

CLICK BUTTON (E10)

- All conditions click the correct button once it has been found

In the no-label condition, a rule called ARB-NL-RETRIEVE-LABEL-FAIL fires that places a chunk called blank in the locate goal. The blank chunk does not have to be retrieved from memory, under the assumption that the participants did not need to explicitly retrieve the fact that the currently needed button had no label, i.e., because none of the buttons had labels.

In the arbitrary condition, a rule called ARB-RETRIEVE-LABEL fires that attempts to retrieve a past use of the button that applies the goal color and also tries to retrieve the chunk representing its label. If this rule succeeds, then the label chunk is placed in a goal slot. If it fails, then ARB-NL-RETRIEVE-LABEL-FAIL fires, placing the blank chunk in the goal slot. In order for this retrieval to be successful, the combined base-level activation and associative strength from the goal-color chunk must exceed the retrieval threshold. As can be seen in Figure 17, which shows the proportion of trials in which the label was retrieved over blocks for the arbitrary condition, this process occurs somewhat gradually but by the end of the run the model is retrieving the label chunk on 92% of the trials. As will be discussed later, the success of this retrieval has important implications in the evaluation phase.

Determine Location

After the locate goal's label slot has been filled, the three controlled search-groups rely on the same retrieve rule for locating the currently needed button. The other, color-match, group is assumed to be relying on a pre-attentive search and thus not relying on memory retrieval. A descriptive representation of these rules appears in Table 11. As can be seen in this table, the pre-attentive rule sets out to locate an object on the screen with a color that matches the color of the rectangle, pushes a color-

button subgoal to evaluate the object, and directs visual attention and the mouse to the object's location. This rule relies on the ACT-R/PM vision module described above, so it always fires successfully. It is thus not subject to the production rule competition in the search versus retrieve paradigm.

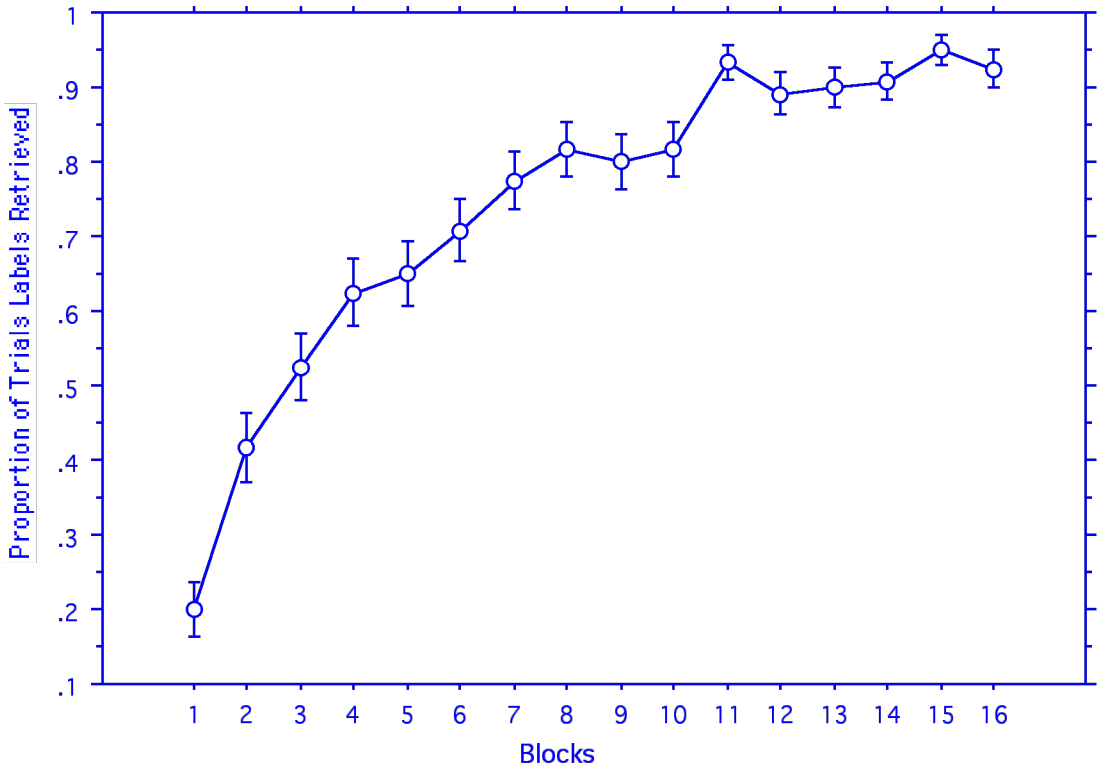


Figure 17. Mean proportion of trials per block in arbitrary condition in which label chunk was successfully retrieved. Error bars represent standard error

The retrieve rule for the controlled search conditions, FIND-BUTTON-RETRIEVE, requires two successful chunk retrievals: one is the chunk encoding a past color-button goal involving the currently needed button and the other is the visual-location chunk that was created by ACT-R/PM in a previous movement of attention to that button (location-chunk). The find-button-retrieve rule is attempted first. If it succeeds, it directs

visual attention and the cursor to the location specified by the location chunk and pushes a new color-button subgoal to evaluate the button at that location. If it fails, then the search rule, FIND-BUTTON-RANDOM-GUESS, is attempted which checks that the retrieve rule has failed enough times (the number of retrieval attempts varies somewhat between conditions - this will be discussed later). If the search rule succeeds, visual attention and the cursor are moved to the location of a randomly-selected button and a color-button subgoal is pushed to evaluate that button. If the search rule fails, then FIND-BUTTON-RETRIEVE-FAIL fires. This rule increments the failed retrieval count and leaves the goal unmodified so that FIND-BUTTON-RETRIEVE rule will try again on the next cycle.

The past color-button goal chunks the model is attempting to retrieve in FIND-BUTTON-RETRIEVE have three slots containing information about the attributes of the button: the *does* slot contains the color applied by the button, the *label* slot contains the button's label, and the *loc* slot contains the button's location. These slots are filled in with the appropriate chunks as the model completes a trial. When a color-button goal is popped (after the correct button has been located), if the button had been encoded previously, then the new button chunk will be identical to a pre-existing chunk and thus, the two chunks will be merged into a single chunk. Therefore, over the course of the model run there will be a single color-button chunk for each of the 12 buttons, enabling base-level activation learning and associative strength learning for these chunks.

There are two key chunks acting as activation sources when the FIND-BUTTON-RETRIEVE rule is attempted: the goal color chunk and the label chunk, if it was retrieved. These chunks are contained in the *loc* and *label* slots of the color-button chunk, so their presence in the goal has the potential to increase the probability that the button chunk

will be retrieved. It turns out, however, that spreading activation can be a double-edged sword due to the fan effect. In the case of the no-label condition, because all of the buttons have the same blank chunk in the label slot, the activation spread to the color-button chunk is less than that for the meaningful or arbitrary conditions, where the label chunk is unique to a given button. This results in a slight decrease in the probability and speed of button retrieval in the no-label condition and is intended to represent the non-distinctiveness of the buttons in this condition.

Table 11. Text descriptions of production rules used for the four conditions in the determine-location sub-phase of task performance. Production rule names are underlined. The ACT-R syntax may be found in Appendix E.

Color-Match (Pop-Out) FIND-BUTTON-PRE-ATTENTIVE IF the goal is to locate the required button and the location of the button is currently unknown and there is a button-percept on screen whose color matches the goal-color (and the motor-module is free) THEN move attention and the cursor to the location of that percept and set a subgoal to evaluate the button at that location
Arbitrary, Meaningful and No-Label (Location-Retrieve) FIND-BUTTON-RETRIEVE IF the goal is to locate the required button and the location of the button is currently unknown and there is a memory trace of using a button that does goal-color and there is a memory trace of the location of that button (and the motor-module is free) THEN move attention and the cursor to the location of that button and set a subgoal to evaluate the button at that location

FIND-BUTTON-RANDOM-GUESS

IF the goal is to locate the required button
and the location of the button is currently unknown
and there is a button-percept on screen
and the required number of retrieval attempts have failed
(and the vision and motor-module are free)
THEN
move attention and the cursor to the location of that percept
and set a subgoal to evaluate the button at that location

FIND-BUTTON-RETRIEVE-FAIL

IF the goal is to locate the required button
and the location of the button is currently unknown
and there is a button-percept on screen
THEN
increment the retrieval attempts count

As previously mentioned, the three controlled search conditions vary somewhat in the number of times the FIND-BUTTON-RETRIEVE rule will be attempted before FIND-BUTTON-RANDOM-GUESS fires. The model assumes that the additional time taken by the no-label and arbitrary groups over the meaningful and color-match groups found in the data (see Figure 13) is being used for additional attempts to retrieve the currently needed color-button chunk and its location. Based on the qualitative properties of Figure 13, the number of attempts was set 1 for the meaningful condition, 2 for the arbitrary condition, and 3 for the no-label condition. The increased number of retrieval attempts in the no-label condition helps the model overcome the decreased probability of successfully firing the FIND-BUTTON-RETRIEVE rule resulting from the large fan from the blank chunk to color-button chunk.

The rectangle study times (i.e., the time spent attending to the colored rectangle at the beginning of a trial) generated by the model are presented in Figure 18. As can be seen in this figure, the controlled search groups all show a gradual decrease in the amount of time spent attending to the center over blocks. This is due to a combined

reduction in the time to retrieve the label, color-button, and location chunks as well as a reduction in the number of retrieval attempts before FIND-BUTTON-RETRIEVE succeeds. The model's fit to the data is quite good at $r^2 = .88$ (see Table 12 for mean absolute deviations by condition).

Approximate-Trials

Once the model has pushed the color-button goal, the next step involves moving visual attention and the cursor to that location. Here the approximate-trials phenomenon is worked into the model. This phenomenon was attributed to either noise in eye movements and/or noise in location memory. This issue was not satisfactorily resolved based on the empirical data, so the model makes no strong commitment as to the true cause. The model simply mimics the effect by adding noise to the location chunk sent to ACT-R/PM's vision module. Without the addition of this noise, the vision module will accurately shift attention to the object at the x and y coordinate contained in the location chunk. The noise added to the location chunk results in attention being shifted to a button adjacent to the correct one on approximately 12% of the trials, just as in the empirical data.

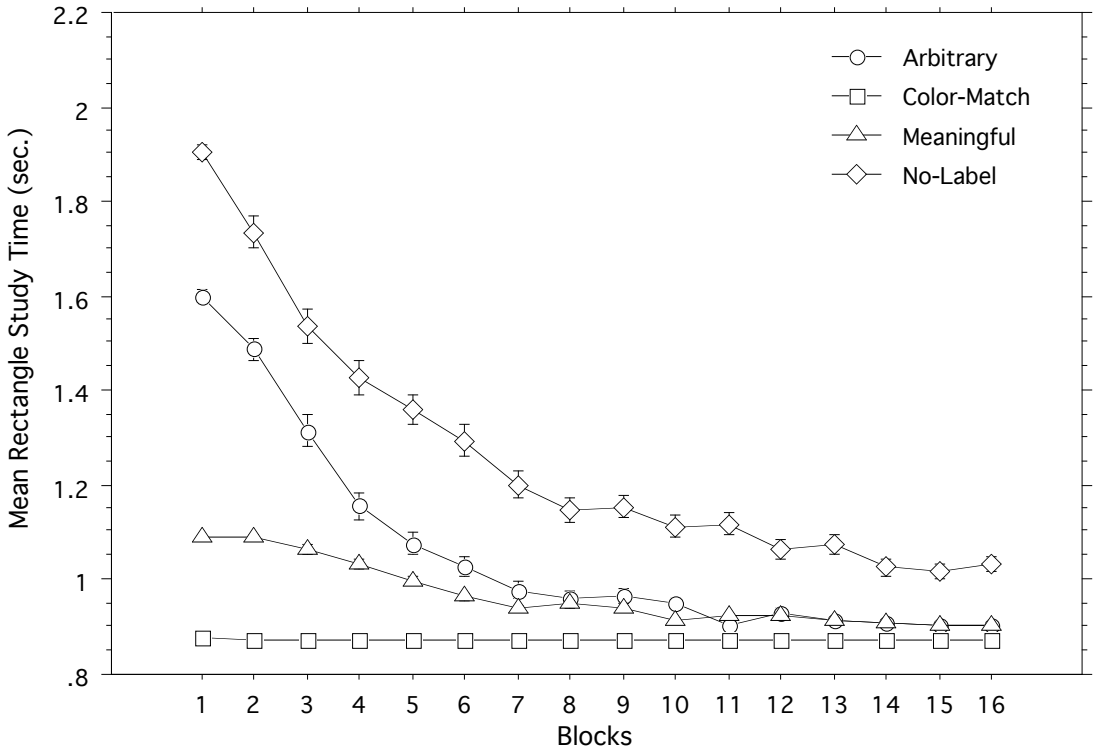


Figure 18. Mean time spent by the model attending to the center rectangle by block and label condition

This noise is added to the commands to move visual attention and the cursor on the action sides of both the FIND-BUTTON-PRE-ATTENTIVE search rule and in the FIND-BUTTON-RETRIEVE rule. The addition of this noise in the FIND-BUTTON-PRE-ATTENTIVE rule may be taken as a tentative assumption of a speed/accuracy trade-off in the eye-movement from the rectangle to the button. The addition of the noise to the controlled-search retrieve rule may be taken as a tentative assumption of imprecise location knowledge as well as a speed/accuracy trade-off in the eye-movement. The effects of these assumptions on the model will be discussed in detail later.

Table 12. Mean absolute deviations of model fit by behavioral measure and condition along with overall r^2 values.

Label Condition	Center Rectangle Time (seconds)	Buttons Evaluated (buttons/trial)	ToolTip Accesses (tips/trial)	Average Evaluation Time (seconds)	Performance Time (seconds)
Color-Match	0.11	0.18	0.00	0.10	0.11
Meaningful	0.06	0.47	0.01	0.06	0.20
Arbitrary	0.19	0.37	0.13	0.19	0.40
No-Label	0.15	0.27	0.43	0.18	0.82
Overall Deviation	0.13	0.32	0.14	0.13	0.38
Overall r^2	.88	.88	.94	.89	.94
	Figure 18	*Figure 19*	*Figure 20*	*Figure 21*	*Figure 22*

Number of Buttons Evaluated

Subject to the noise just described, the ability to successfully retrieve, or in the case of the color-match condition, pre-attentively determine, the location of the currently needed button reduces the number of buttons evaluated over blocks. The data from Experiment II are presented in Figure 14, and Figure 19 shows the results of the model run, which reveals a good fit (r^2 = .88; see Table 12 for mean absolute deviations by condition). As in the data, the number of buttons evaluated in the color-match condition remains steady throughout the blocks at about 1.5 buttons per trial. The pre-attentive search conducted by the model for this condition precludes the need to conduct an extensive search of the screen, thus keeping the value quite low. The value

consistently exceeds a perfect value of 1, however; due to the location noise added to the ACT-R/PM vision module.

The controlled-search conditions start off at an average of 6.3 buttons per trial in block 1, slightly higher than in the data, and gradually decrease to about 1.5 buttons per trial by block 10. This effect is due to the model eventually attaining the ability to retrieve the required button and location chunks in the search phase. Successful retrieval, in turn, enables the retrieve rule to fire instead of the random search rule, thus decreasing the number of buttons to which the model had to shift attention on a given trial. As in the data, these three label conditions show a similar rate of decrease.

Evaluation Phase

Encode-Label

For all conditions, the model begins the evaluation phase by encoding the label on the currently attended button and placing the label-chunk in the color-button goal. For the arbitrary and meaningful conditions, this rule is ENCODE-LABEL, for the no-label condition it is NL-ENCODE-LABEL, and for the color match condition it is COLOR-ENCODE-LABEL. In the color-match condition, the chunk is placed in the *label* slot and in the other conditions it is placed in the *crnt-label* slot. The model then sets out to determine whether or not the currently attended button is the one currently needed. Descriptive versions of the rules for how the four label conditions proceed from here are presented in Tables 13, 14 and 15.

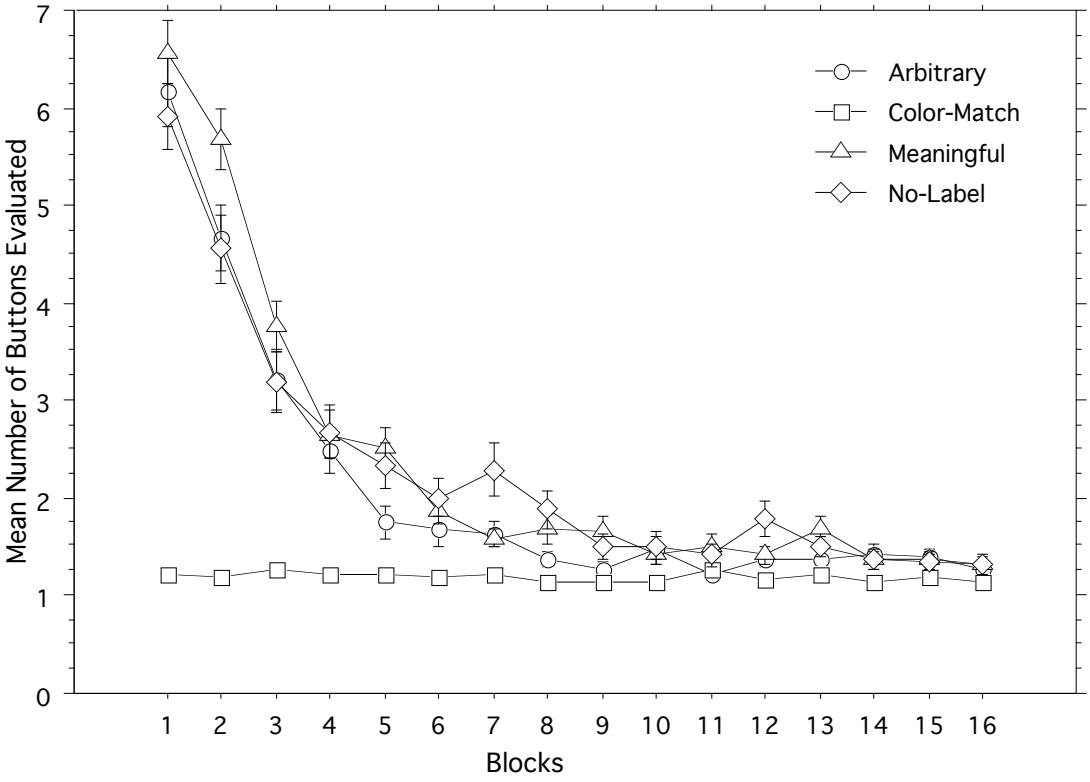


Figure 19. Mean number of buttons evaluated by the model by label condition and blocks.

Compare

For the color-match condition, evaluation simply entails comparing the chunk encoding the rectangle color, stored in the *does* slot of the current color-button goal, to the color-chunk just placed in the label slot (see Table 13). If they are the same, then COLOR-LABEL-CORRECT fires, putting the button's location in the *loc* slot of the goal and popping the goal. If the slots contain different chunks, then COLOR-LABEL-WRONG fires, setting off another pre-attentive search for the goal-color (identical to the one conducted in the search phase).

Table 13. Text descriptions of production rules used for the color-match condition in the evaluation phase of task performance. Strategy is in parentheses and production rule names are underlined. The ACT-R syntax may be found in Appendix E, section 8.

Color-Match (Label-Matching)

COLOR-LABEL-CORRECT

IF there is a goal to evaluate the current button
and the label of the currently-attended button is the same color as the goal color
THEN
put the location of the button in the goal
pop the goal

COLOR-LABEL-WRONG

IF there is a goal to evaluate the current button
and the label of the currently-attended button is not the same color as the goal color
and there is a button-percept on screen whose color matches the goal-color
(and the vision and motor-module are free)
THEN
move attention and the cursor to the location of that percept
pop the goal

The process for the meaningful text condition is similar to the process for the color-match condition in that it simply requires a comparison of objects stored in the goal (see Table 14). In this case, however, the comparison is between the label-chunk stored in the label slot (retrieved in search phase) and the label-chunk in the *crnt-label* slot. If it is the same chunk, MN-LABEL-CORRECT will fire, putting the location of the button in the goal's *loc* slot, and popping the goal. If the slots contain different chunks, then on about two thirds of the trials, the MN-LABEL-WRONG rule will fire, which sets up the goal to try another button. On the other one-third of the trials, the MN-LABEL-WRONG-ENCODE rule fires, meaning that the model will create a color-button chunk representing the currently attended button prior to modifying the goal to try again.

The chunk created by MN-LABEL-WRONG-ENCODE is identical to the chunk that would have been created had current button been the correct one, i.e., its slots contain the

color associated with the button, its label and its location. In this manner, color-button chunks for buttons that had not yet been clicked will exist in declarative memory. This arrangement is intended to model participants' encoding of buttons encountered during their search for the currently needed button. If the model did not encode color-button chunks for non-clicked buttons, then the FIND-BUTTON-RETRIEVE production would never fire successfully in the first block of trials (which includes only one use of each of the twelve buttons). If the model always encoded non-clicked buttons, then learning would tend to occur more quickly than was shown in the data.

Table 14. Text descriptions of production rules used for the meaningful condition in the evaluation phase of task performance. Strategy is in parentheses and production rule names are underlined. The ACT-R syntax may be found in Appendix E.

Meaningful (Label-Matching)

MN-LABEL-CORRECT

IF there is a goal to evaluate the current button
and the label on the currently-attended button matches the label in the goal
THEN
put the location of the currently-attended button in the goal
pop the goal

MN-LABEL-WRONG

IF there is a goal to evaluate the current button
and the label on the currently-attended button does not match the label in the goal
THEN
set the *loc* slot of the goal to try another button

MN-LABEL-WRONG-ENCODE

IF there is a goal to evaluate the current button
and the label on the currently-attended button does not match the label in the goal
THEN
encode a memory of having seen the button that does the color associated with the current label
set the *loc* slot of the goal to try another button

In the arbitrary condition, the model chooses a different evaluation strategy depending on whether the label was successfully retrieved during the search phase (see Table 15). If it was, then ARB-LABEL-CORRECT and ARB-LABEL-WRONG will be in competition with one another. These rules correspond to the label-matching strategy described above for the meaningful group, i.e., the label chunk in the *label* slot of the color-button goal is compared to the label chunk in the *crnt-label* slot. If the labels match, then the location is placed in the *loc* slot and the goal pops. If the labels do not match then the goal is modified to try another button.

If the arbitrary label was not retrieved during the search phase, then the rules considered by the arbitrary and no-label conditions are identical. There are two location-recognition rules which involve recognizing whether the currently-attended button is the correct one, ARB-NL-RECOGNIZE-BUTTON-OK and ARB-NL-RECOGNIZE-BUTTON-WRONG, as well as a rule that pushes a subgoal to wait for a ToolTip when the previous rules fail, ARB-NL-RECOGNIZE-FAIL-GET-TIP. In the first two location-recognition rules, determining whether the current button is the correct one requires retrieving both the color-button chunk representing a past use of the button and the location chunk representing the button's location. The requirement that the location chunk be retrieved is intended to represent the claim that the participants in the arbitrary and no-label groups were relying on location information in the evaluation phase. The underlying assumption is that the retrieval of the location chunk represents the use of location as a criterion for evaluating the button.

Performance differences between the arbitrary and no-label conditions in the evaluation phase emerge from two sources. First, in the arbitrary condition, the model

can sometimes rely on the faster, label-matching, strategy, whereas in the no-label condition the must either rely on the location-recognition strategy or a ToolTip. Second, even when attempting the location-recognition strategy, the arbitrary condition has a higher probability of retrieving the button and location chunks than the no-label condition due the larger fan of the blank label chunk in the no-label condition. As such, the location-recognition rules fail more often in the no-label condition resulting in more ToolTip requests (see Figure 20). The number of tips accessed by the model correspond well with the data from the participants (see Figure 12), with an $r^2 = .94$ (see Table 12 for mean absolute deviations by condition).

Table 15. Text descriptions of production rules used for the arbitrary and no-label conditions in the evaluation phase of task performance. Strategies are in parentheses and production rule names are underlined. The ACT-R syntax may be found in Appendix E.

Arbitrary (Label-Matching) ARB-LABEL-CORRECT IF there is a goal to evaluate the current button and the label on the currently-attended button matches the label in the goal THEN put the location of the currently-attended button in the goal pop the goal ARB-LABEL-WRONG IF there is a goal to evaluate the current button and the label on the currently-attended button does not match the label in the goal THEN set the *loc* slot of the goal to try another button
No-Label and Arbitrary (Location-Recognition) ARB-NL-RECOGNIZE-BUTTON-OK IF there is a goal to evaluate the current button and there is a memory trace of using the currently-attended button which indicates that the button applies the goal color and there is a memory trace of the location of that button THEN

put the location of the currently-attended button in the goal
pop the goal

ARB-NL-RECOGNIZE-BUTTON-WRONG

IF there is a goal to evaluate the current button
and there is a memory trace of using the currently-attended button which indicates that the button does not apply the goal color
and there is a memory trace of the location of that button
THEN
set the *loc* slot of the goal to try another button

ARB-NL-RECOGNIZE-FAIL-GET-TIP

IF there is a goal to evaluate the current button
and attempts to recognize what the button does have failed
THEN
set a subgoal to wait for a ToolTip on the button

When a ToolTip reveals that the current button is not the correct one, on half of the trials the model encodes a color-button chunk containing information about the current button. This is analogous to the behavior of the MN-LABEL-WRONG-ENCODE rule discussed above. For the remaining trials, the model modifies the current color-button goal so that the model will then search for another button to evaluate.

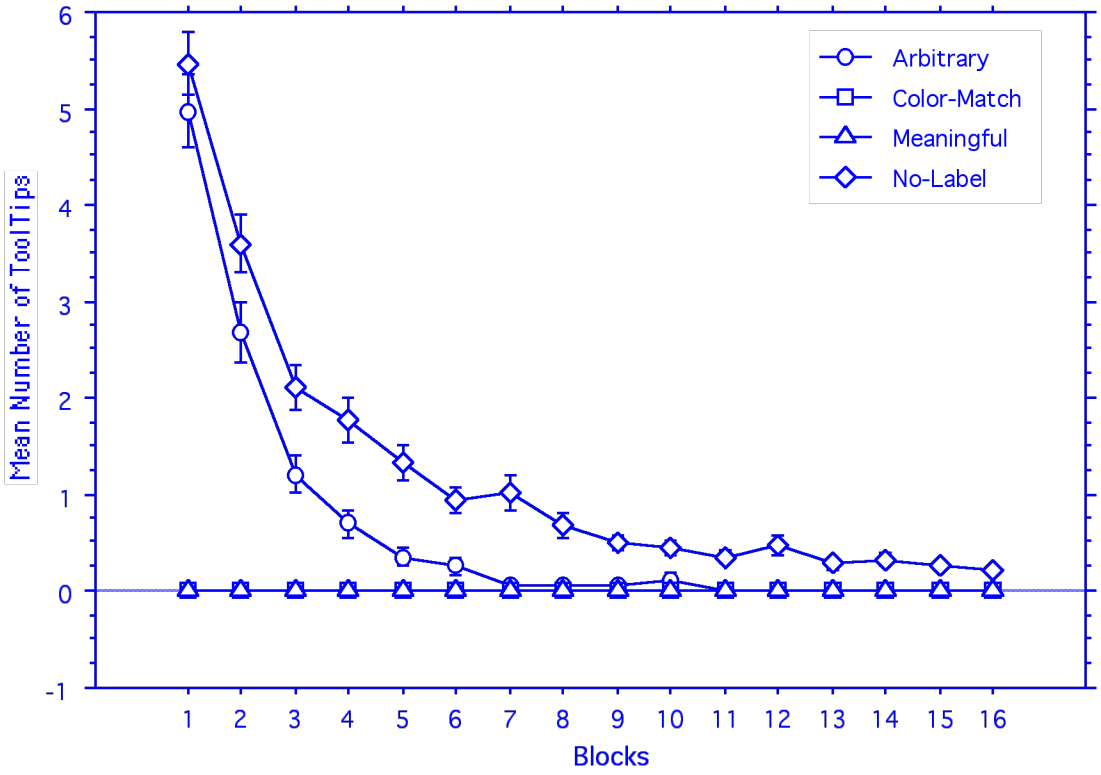


Figure 20. Mean number of ToolTips accessed by the model by label condition and blocks.

When the model sets out on this search, there are two rules in competition with one another. In the SEARCH-NEXT-BUTTON-RETRIEVE rule, the model attempts to retrieve the correct color-button button and location chunks (akin to FIND-BUTTON-RETRIEVE) and, if successful, moves attention to the retrieved location. In the SEARCH-NEXT-BUTTON rule, the model simply moves attention to the closest button it has not yet evaluated. As such, the model is not locked into a visual search for an entire trial if FIND-BUTTON-RETRIEVE fails to fire in the search phase. The presence of the SEARCH-NEXT-BUTTON-RETRIEVE rule also enables the model to recover from approximate-trials relatively quickly. In these trials, FIND-BUTTON-RETRIEVE fired successfully, but due to the noise added to command to move

attention, the model ended up attending to an incorrect button. Because the required retrievals are identical in FIND-BUTTON-RETRIEVE and SEARCH-NEXT-BUTTON-RETRIEVE, the model will have a high probability of moving attention to the correct button if SEARCH-NEXT-BUTTON-RETRIEVE is attempted.

Average Evaluation Time

Given the assumptions and rules just described, how well does the model capture the differences between label conditions in the evaluation phase? The data for average button evaluation time is plotted in Figure 11, and the results from the model are shown in Figure 21. The model captures the basic trends in the data, resulting in a good fit (r^2 = .89; see Table 12 for mean absolute deviations by condition). In both figures, the arbitrary conditions improve much faster than the no-label conditions, due largely to the decreased reliance on ToolTips. In the model, this is also due to the fact that the label-matching rules for the arbitrary group (ARB-LABEL-CORRECT and ARB-LABEL-WRONG) are faster than the location-recognition rules (ARB-NL-RECOGNIZE-BUTTON-OK and ARB-NL-RECOGNIZE-BUTTON-WRONG) because the latter rules require explicit memory retrievals, whereas the former do not.

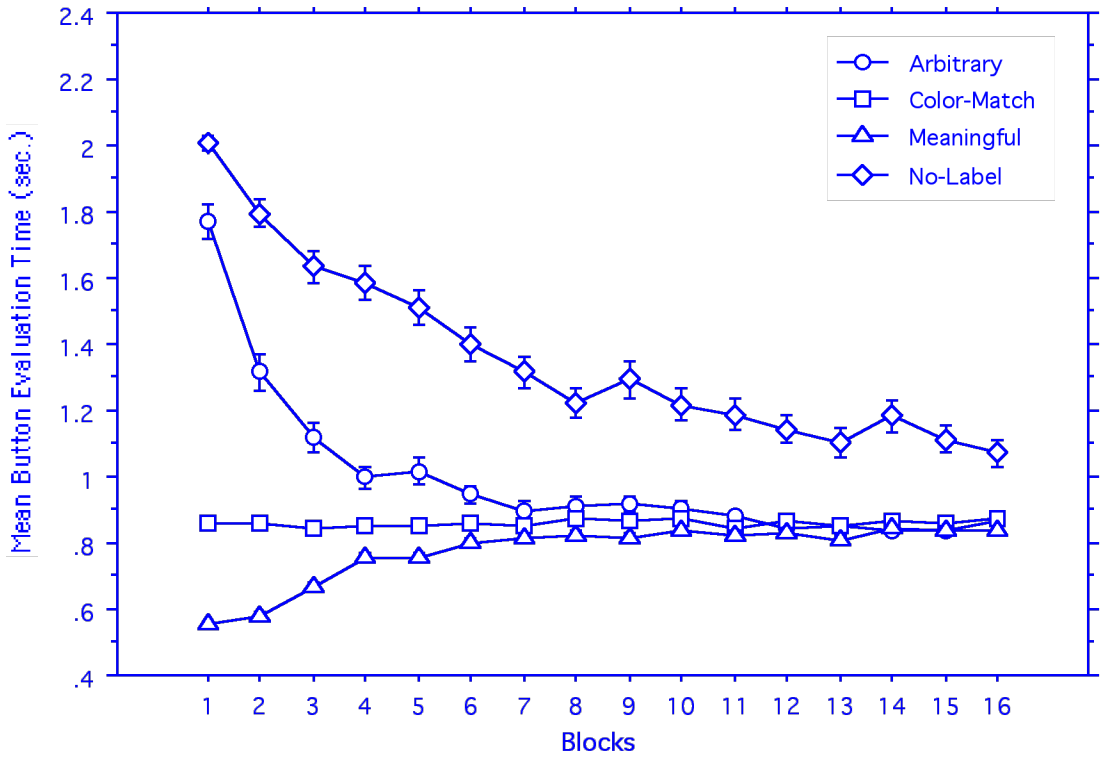


Figure 21. Mean time the model spent attending to each button by label condition and blocks.

Performance Time

Thus far, it has been shown that the model provides good fits to various fine-grained components of performance in Experiment II, such as the number of buttons evaluated per trial and average evaluation time. The next question is, putting all of the component parts of the model together, how well does the model account for the overall performance times? Figure 22 shows the trial times generated by the model and Figure 10 shows the data. A comparison of these figures reveals that the model does a nice job of capturing the learning trends and the relative differences between the groups. The fit is quite good at $r^2 = .94$ (see Table 12 for mean absolute deviations by condition). Thus,

the model captures the performance data at both a coarse and relatively fine grain size of analysis.

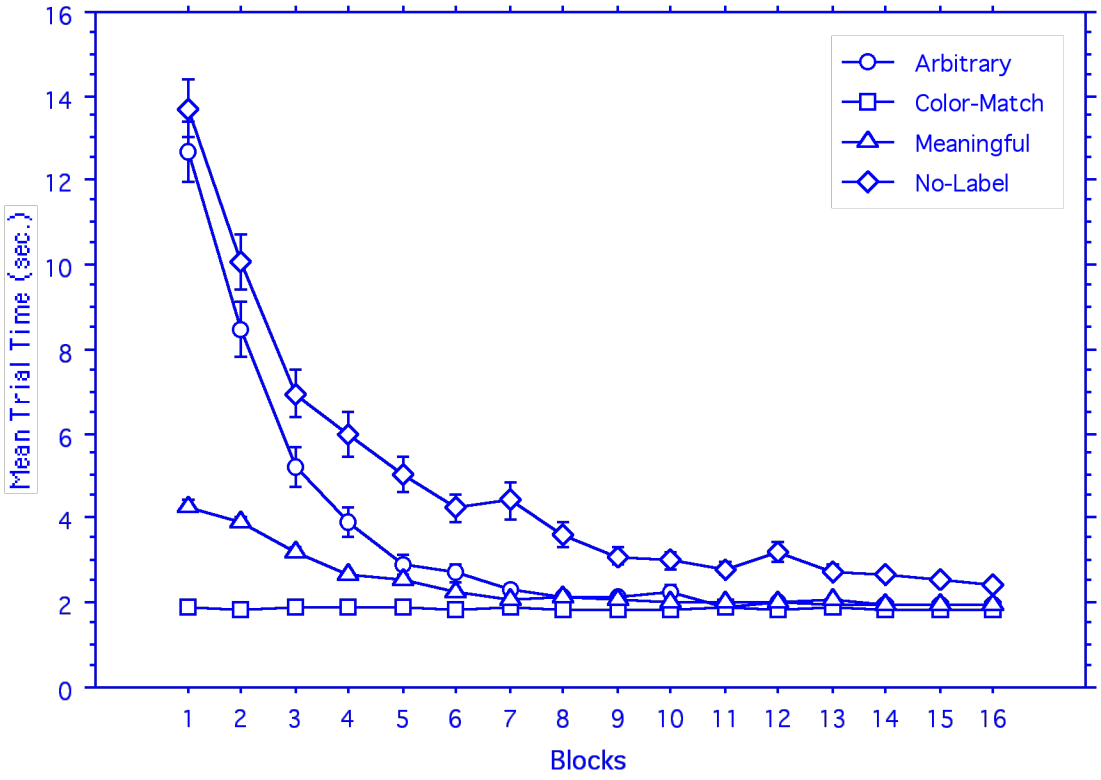


Figure 22. Mean trial times generated by the model by label condition and blocks.

Location Knowledge

Using the activation-levels of various chunks in the models' memory at the end of the run and some assumptions about the retrievals required for performance on the location memory test given at the end of Experiment II, it is possible to derive gross predictions about the model's performance on this test. Because the location memory test requires that the locations be reconstructed without visual access to the buttons, it

is assumed that accurate performance requires, at minimum, the ability to retrieve both the required color-button chunk and its associated location chunk.

The approach taken in this analysis is to calculate the joint probabilities of retrieving the button and location chunks for each condition using chunk activation levels at the end of the experiment, and then to compare those probabilities to the location memory test scores. As a caveat, it is important to point out that because the performance predictions of the model are based solely on the probability of retrieving the chunks assumed to be required for successful performance, and the actual performance of this test presumably goes beyond the mere ability to retrieve this knowledge, the predictions generated in this analysis are expected to only generally reflect test performance.

The probabilities of retrieving the button and location chunks were separately calculated via the formula below (Anderson & Lebiere, 1998, p. 74) and multiplied to get the joint probability of retrieving them both on a given cycle.

$$\text{Probability of retrieval} = 1 / 1 + e^{-((A - \tau) / s))} \quad (1)$$

In this equation, τ is the activation retrieval threshold, which, as shown in Table 9, was set at 2.3, *s* is the activation noise, set at 0.7, and A is the activation of the chunks at the end of block 16. The chunk activations were calculated assuming that the appropriate color chunk and label chunk were activation sources. When ACT-R attempts to retrieve chunks, each activation source spreads an equal proportion of its inter-associative strength (ias) to the chunk being retrieved. In the formula below, there are two activation sources, so each spreads half of its ias to the chunk being retrieved.

$$A = \text{chunk base-level} + (0.5 \times \text{color}\rightarrow\text{chunk ias}) + (0.5 \times \text{label}\rightarrow\text{chunk ias}) \quad (2)$$

The joint retrieval probabilities generated from the model are presented in Table 16 along with the error scores from the location memory test. Because the error scores are the average deviation in pixels, these scores map inversely to the retrieval probabilities, such that higher retrieval probabilities should result in lower error scores. As can be seen in this table, the retrieval probabilities capture the major trend in the data, namely, the performance difference between the controlled search and color-match groups, resulting in a good fit, $r^2 = .90$.

Table 16. Probability of retrieving the button-chunks and location-chunks by condition along with mean absolute deviation scores (in pixels) from the Experiment II location memory test.

Label Condition	$P_{Button\&Location}$	Location Memory Test Score
Color-Match	0.33	99.29
Meaningful	0.93	45.86
Arbitrary	0.95	29.43
No-Label	0.75	41.51

This trend is captured via the assumption that the controlled search interfaces require retrieval of both the button and location chunks during task performance, whereas the color-match interface did not. This results in a higher probability of successfully retrieving these chunks at the end of the run in the controlled-search conditions than in the color-match condition. As the model completes the task in the color-match condition, the color-button and location chunks only receive base-level activation boost by virtue of ACT-R's chunk merging mechanism. Because the controlled

search groups were also retrieving these chunks: (1) the chunks' base-level activations increased at a faster rate, and (2) interassociative strength built up between the activation sources and the chunks being retrieved, resulting in a higher probability of retrieval.

Experiment I Performance Disruption

The above approach was also used to evaluate the extent to which the model could account for the performance disruption effects from Experiment I. Specifically, the fit to the Location Knowledge Scores was assessed. Location knowledge score is the proportion of Directs and Verifies in the assessment blocks; i.e., trials in which the participants located and clicked the correct button, and in doing so either accessed no tips (Direct) or accessed a ToolTip on only the correct button (Verify). Given the structure of the FIND-BUTTON-RETRIEVE rule in Table 11, to locate the correct button, the participants would have had to retrieve both the correct color-button chunk and location chunk. On failure to retrieve these chunks, the model would conduct a random search of the screen. Thus, it is assumed that the probability of retrieving these chunks can be used to predict the model's ability to recover if the labels were suddenly removed.

The activation values used to calculate the retrieval probabilities were calculated as they were above (see Equation 2) and were sampled at trials 90 and 450 from the model run. The activation level sampled at trial 90 corresponds to the early assessment from Experiment I (5 blocks of experience) and the activation sampled at trial 450 corresponds to the late assessment (25 blocks of experience).

The retrieval probabilities were calculated in a slightly different manner than for the location memory test analysis. As indicated in the description of the model, the no-

label condition tries the FIND-BUTTON-RETRIEVE rule three times before just guessing a location, thus increasing the joint probability of retrieving the button and location chunks somewhat. When the labels were removed from the buttons in Experiment I, all of the groups were faced with an interface identical to that of the no-label group, effectively placing them all in the no-label condition after that point. Thus, in calculating the retrieval probabilities from the model, it was assumed that all of the conditions would, like the no-label condition, attempt to retrieve the button and location chunks (i.e., attempt find-button-retrieve) three times before guessing a location. The following formula for calculating the joint probabilities therefore assumes three retrieval attempts.

$$P(\text{button \& location}) = 1 - (1 - (P(\text{color-button}) \times P(\text{location})))^3 \quad (3)$$

Table 17 shows the location knowledge scores from the early and late assessment blocks by group for Experiment I and the associated retrieval probabilities generated from the model. In general, the probabilities from the model map reasonably well to the data, with an $r^2 = .81$. The model captures the major trends in the data, with the color-match group predicted to perform worse than the controlled search groups in both early and late assessment, and with all groups predicted to perform better in the late assessment than the early assessment.

Table 17. Probability of retrieving the button-and location-chunks by condition mapped to the Location Knowledge Score from Experiment I assessment trials. Predictions generated from the model are in parentheses.

Assessment Time	Label Condition	Location Knowledge Score
Early	Color-Match	0.52 (0.41)
	Meaningful	0.78 (0.97)

	Arbitrary	0.87 (0.98)
	No-Label	0.82 (0.87)
Late	Color-Match	0.79 (0.91)
	Meaningful	0.96 (0.99)
	Arbitrary	0.99 (0.99)
	No-Label	0.97 (0.99)

The error in prediction comes largely from the model's under-prediction of the performance of the color-match group and over-prediction of the meaningful and arbitrary groups in the early assessment block. In the model, the activation levels for the color-match condition increase solely due to the increases in base-level activation that result from merging of identical chunks; i.e., because the button and location chunks are never being retrieved, no source spread develops from the label or color-chunks. Thus, the activation levels, and therefore the retrieval probabilities, for this condition increase at a much more gradual rate than for the other groups. The rate of activation increase for the model is more gradual than in the data. The model also slightly over-predicts the performance of the arbitrary and meaningful text groups in the early assessment. Unlike the color-match condition, the activation levels and thus retrieval probabilities for the button and location chunks in these conditions are increasing too quickly.

Summary

Overall, the model provides a compelling account of the critical components of the data from both experiments. The model captures the trends in the eye and

performance data from Experiment II, with fits ranging from .88 to .94 (see Table 12). Although performance on the location memory test was not modeled, an analysis of the retrieval probabilities for button-chunks and location-chunks for each of the label conditions revealed that the model could capture relative differences in test scores between groups. Likewise, the retrieval probabilities were also able to provide an account of performance disruption shown by the groups in Experiment I.

An underlying assumption of the model is that participants behaved rationally, seeking maximum gain at minimum cost. As such, participants in the various conditions adopted the least-effort strategies allowed by their interfaces, yet were shown to be equally fast with enough practice. The least effort strategies chosen by participants tended to be display-based; i.e., they enabled participants to rely on perceptually available information rather than memory retrievals. As summarized in Table 18 and discussed in more detail below, the model assumes that participants only adopted a memory-intensive means of interaction when a display-based strategy was not available, such as in the case of the controlled search groups in the search phase, or the arbitrary and no-label groups in the evaluation phase.

Table 18. Assumed strategy use by task phase and label condition. Display-based strategies are in *italics*.

Label Condition	Search Phase	Evaluation Phase
Color-Match	*Pop-Out*	*Label-Matching*
Meaningful	Location-Retrieval	*Label-Matching*
Arbitrary	Location-Retrieval	*Label-Matching*/ Location-Recognition

No-Label	Location-Retrieval	Location-Recognition

The lowest cost, color-match, condition (see Table 1) provided participants with the opportunity to rely on the pre-attentive, pop-out effect, a fast and accurate means of locating and evaluating the currently-needed button. The data showed that participants attended to very few buttons per trial and showed no significant improvement in performance over blocks, suggesting that they took the opportunity to rely on the informative perceptual cues in the interface. The model captured this behavior by relying on a pre-attentive search to locate the currently needed button in the search phase and using a label-matching strategy in the evaluation phase.

When forced to rely on their location knowledge in the assessment blocks of Experiment I or in the location memory test from Experiment II, participants in the color-match condition were shown to perform more poorly than participants in other conditions. The ability of the model to capture the performance on these tests relied on ACT-R's activation learning mechanisms. Because the production rule implementation of the display-based strategies assumed to be adopted by the color-match participants do not require memory retrievals, the base-level activations and associative strengths for the chunks required by the production rule implementations of the memory-intensive strategies did not increase at as fast a rate as they did for the other groups. Thus, when these button and location chunks needed to be retrieved, their relatively low activation levels resulted in a lower probability of retrieval than the other groups. The probability of retrieval of these chunks does increase as the model completes trials, however; and is

thus consistent with experimental data indicating location knowledge increased as a function of experience.

The model also assumes that the controlled search groups relied on least effort strategies when possible. In the search phase, these groups learned and used the locations of the buttons to direct visual attention rather than continuing to rely on the more time consuming visual search of the screen. This was manifested in a decrease from (almost) chance performance in the number of buttons evaluated per trial in block 1 to a single button per trial in later blocks. The model exhibits this behavior through its preference for the FIND-BUTTON-RETRIEVE rule over the FIND-BUTTON-RANDOM-GUESS RULE, such that it attempts to retrieve the location of the currently needed button prior to conducting a search.

In the evaluation phase, the meaningful group spent a small and constant amount of time per button. The model accounts for this by assuming that this group had prior knowledge of the relationship between colors and the words describing them and was using this knowledge in a display-based, label-matching strategy. The model can always retrieve and compare the target label to the labels on the buttons being evaluated.

Because no prior relationship existed between the colors and the icons used in the arbitrary condition, participants in this group could not use the label match strategy unless they had learned the association between the colors and icons. The accuracy score from the icon memory test given to the arbitrary group in Experiment I was quite high, at 81%, indicating that participants had indeed learned these associations. In the arbitrary condition, the model attempts and eventually succeeds in retrieving the target label, thus allowing it to use the display-based label-match strategy. Prior to successful

retrieval of the target label, the model attempts the more effortful and time-consuming location-recognition rules (ARB-NL-RECOGNIZE-BUTTON-OK and ARB-NL-RECOGNIZE-BUTTON-WRONG), which often fail early in the run, thus resulting in reliance on a ToolTip for evaluation. As in the data, the average button evaluation times for the model (see Figure 21) start off roughly equivalent to the no-label group in block 1, but show a comparatively rapid decrease due to successful label retrievals and use of the label-match strategy.

Participants in the no-label condition exhibited a much more gradual decrease in average button evaluation time as compared to the arbitrary group. In the model, this is due to sustained reliance on the location-recognition rules (ARB-NL-RECOGNIZE-BUTTON-OK and ARB-NL-RECOGNIZE-BUTTON-WRONG), which have a decreased probability of successfully matching due to the large fan from the blank label chunk to the 12 color-button chunks. The non-distinctiveness of the buttons' visual appearance decreases the probability of being able to retrieve the needed color-button chunk, and thus the location-recognition rules fail, leading to more lingering reliance on ToolTips (see Figure 20) and increased button evaluation time (see Figure 21).

Limitations of the Model

The model has little to say about the inherent cause of the disproportionate number of approximate-trials. As previously stated, the model only addresses this effect to the extent that it mimics the behavioral effects. This is accomplished by adding noise to the location chunk passed to ACT-R/PM's vision module. The effect of excluding this noise from the model run is essentially to decrease the number of buttons evaluated and

the trial times. For example, the model running under the color-match condition will always only attend to one button. Likewise, once the model can retrieve the button-chunks and location-chunks in the controlled search conditions, it will always move visual attention directly to the correct button. From a performance perspective, the addition of noise improves the fits to the data by preventing it from systematically underestimating the relevant values in the data. In mapping this assumption to human behavior, the absence of noise would correspond to consistent and perfectly accurate saccades to the precise locations of intended targets.

Another potential limitation of the model is that it does not include any strategic intention to learn the locations of items; the model engages in an incidental type of learning. It is possible that participants could have undertaken to learn the locations of items in order to improve performance via some rehearsal strategy. Indeed, a few participants admitted to using rehearsal and other, more complicated, strategies for encoding the associations between labels, colors, buttons and locations. It is important to point out, however, that any intentional rehearsal or retrieval strategy would tap the same ACT-R sub-symbolic declarative learning mechanisms relied upon by the model, i.e., the base-level and associative learning mechanisms.

Theoretical Implications of the Model

So what does the model have to say about the question of whether spatial information is encoded automatically? As described in the literature review, studies examining this claim typically involve single presentations of stimuli followed by a recognition or recall test (Lansdale, 1991; Naveh-Benjamin, 1987; Naveh-Benjamin,

1988). When recognition or recall is poor, this is taken as evidence against automatic encoding. What these paradigms seem to be tapping is the ability to recognize or retrieve the required information from memory, not whether the information was actually encoded. This distinction is made clear in the model. Because the ACT-R/PM vision module encodes a visual-location chunk for each location attended by the model, a potentially retrievable knowledge structure representing that location will exist in declarative memory. Thus, encoding of location is not deliberate but rather a by-product of visual attention. Hence the ACT-R architecture serves to instantiate a particular sense of the phrase “automatic encoding”. If the model has attended to the location only a few times, however; the probability of retrieving the location chunk is quite low due to a sub-threshold activation level. Thus, the subsequent retrieval of encoded location knowledge is not guaranteed but rather probabilistic.

Multiple visits to a particular location result in a boost of the base-level activation of its chunk, and therefore in an increase in the probability of the chunk's subsequent retrieval. This predicts that the probability of successfully retrieving a location chunk will increase as a by-product of having attended to that location. The probability of retrieval will increase more quickly in tasks which require repeated shifts of attention to objects that remain in constant locations. Further, these probabilities will increase irrespective of explicit intent to learn locations, although strategies such as rehearsal would be expected to speed up the learning process, i.e., by increasing base-level activations and associative strength.

The model's account of location learning does not require the addition of any new perceptual or cognitive mechanisms to ACT-R/PM; it relies on the same ACT-R

mechanisms that underlie the learning of typical declarative knowledge, such as arithmetic facts, (e.g., Lebiere & Anderson, 1998). Once encoded by the vision module and placed into declarative memory, location knowledge is no different from other declarative knowledge: it is subject to power-law decay, increases in activation, and associative strength learning. These mechanisms, used to provide accounts of performance on various other tasks, were able to provide a compelling account of this previously unmodeled phenomenon. Thus, the model inherited the constraints as well as the previous successes of these mechanisms from previous modeling efforts. In turn, the success of this modeling effort provides additional support for their explanatory power.

At a general level, three primary implications for a theory of location learning emerge from this modeling effort: (1) locations are encoded as a by-product of attention, (2) once encoded in memory, location knowledge is subject to the same mechanisms as other declarative knowledge, such as associative learning and decay, such that, (3) the ability to retrieve location knowledge, like other knowledge (e.g., a phone number), requires repetition, practice, or explicit rehearsal. It is important to note that these implications are not specific to the structure of the model, per se, but rather emerge from the default behavior of the vision module and the declarative learning mechanisms of ACT-R/PM.

CONCLUSIONS

The locations of screen objects are central and necessary components to direct manipulation; screen objects must be located, pointed at and clicked on. Knowledge of the locations of sought-after objects can significantly reduce the visual search space, and thus reduce performance times. For this to occur, location knowledge must be retrieved from memory and used to direct visual attention. The pace of the process leading to successful retrieval of location knowledge was shown to be a function of the search cost of using the interface. When the interface provided participants with an opportunity to rely on a low-cost, display-based strategy during the search phase of task performance (i.e., the pop-out effect), then participants chose this display-based strategy over one which required retrieval of location knowledge from memory (i.e., controlled search). Participants in this group were shown to learn locations at a slower rate than in the controlled-search conditions.

The level of reliance on location knowledge during the evaluation phase of task performance was also shown to be a function of interface cost. When the interface supported a display-based strategy in the evaluation phase (i.e., label-matching), participants chose this strategy over one that required retrieval of location knowledge (i.e., location-recognition). Evidence for strategy differences between the low (meaningful) and moderate (arbitrary) evaluation cost groups is provided by the large difference between these groups in the ratio of directs (i.e., assessment block trials in which participants' only action was to click the correct button) to verifies (i.e.,

assessment block trials in which participants accessed a ToolTip prior to clicking the correct button) in Experiment I. Although participants in these groups knew the button locations equally well, as evidenced by their equivalent location knowledge scores, participants in the meaningful group were unable to successfully use the location-recognition strategy and thus had to access a ToolTip, resulting in a verify trial instead of a direct trial.

Participants in the arbitrary group could successfully use the location-recognition strategy (resulting in Directs instead of Verifies) in the Experiment I assessment blocks because they had already been using this strategy. The meaningless labels prevented these participants from label-matching until the color-to-icon associations had been learned. Thus, by the time the labels were removed, participants had already acquired the ability to recognize the button based on its location. The results suggest that the arbitrary group did not rely on the location-recognition strategy throughout task performance, thus providing evidence for strategy differences between the moderate (arbitrary) and high (no-label) evaluation cost groups. The relatively high scores on the Experiment I icon memory test indicated that participants in the arbitrary condition did eventually learn the color to icon associations. Combined with participant responses to this effect in the experiment debriefs and the significantly faster improvement in average button evaluation time for the arbitrary group over the no-label group (who could not label-match), the results suggest that the arbitrary group did eventually use the display-based, label-matching strategy.

Guided by the constraints and learning mechanisms contained in ACT-R/PM and the strategy assumptions outlined above, the Experiment II data were modeled. The

model provided a compelling account of key attributes of participants' behavior, from fine-grained components of interaction such as eye and mouse movements to higher order measures such as the decreased performance time resulting from location learning. The explanatory power of the model was also able to extend beyond the Experiment II task performance data, providing plausible account of the location memory test scores from Experiment II and the performance disruption scores from the assessment blocks of Experiment I.

Theoretical Implications

When analyzed with an eye toward a more general theory of location learning, the model produces three main implications: (1) locations are encoded as a by-product of attention, (2) once encoded in memory, location knowledge is subject to the same mechanisms as other declarative knowledge, such as associative learning and decay, such that, (3) the ability to retrieve location knowledge, like other knowledge, requires repetition, practice, or explicit rehearsal. The implications rely on a distinction between the encoding of locations, which is assumed to occur as a by-product of attentional shifts, and the subsequent retrieval of the location knowledge, which is assumed to be subject to the same constraints and learning mechanisms as other forms of knowledge.

This encoding/retrieval distinction has implications for previous research investigating the question of whether spatial encoding is automatic (e.g., Naveh-Benjamin, 1987; Naveh-Benjamin, 1988). The location memory tests used in this research implicitly required recall or recognition of locations. As such, these tests were not measuring spatial encoding directly, but rather the ability of participants to later

retrieve that knowledge from memory. Given that participants had only seen the stimuli once, and for a relatively short period of time, the model presented above would predict that although locations had been encoded (assuming participants attended to the objects), retrieval performance would be poor due to sub-threshold activation levels..

These theoretical implications originate largely from the ACT-R/PM cognitive architecture, rather than from the structure the model itself. The model relies only on mechanisms currently included in ACT-R in providing its account -- no new learning mechanisms were required. The inheritance of established mechanisms as theoretical constraints is a primary benefit to modeling in a cognitive architecture such as ACT-R.

Implications for Interface Design

The results of this research effort not only provide an empirical and theoretical basis for the positional constancy design guideline, but also demonstrate and explain the performance advantages associated with using distinct and/or representative labels on objects. The empirical data collected in Experiments I and II provided strong evidence that people learn locations and can use this location knowledge to improve task performance. The theoretical account of the data provided by the model implies that location learning occurs as a by-product of interaction such that, without specific intent to do so, users gradually learn the locations of the interface objects to which they attend. However, in order for the mechanisms underlying this learning to accomplish their task, the object locations must remain constant.

Based on the behavior of the model, variable object locations would hamper location learning in two key ways. First, the base-level activation boost resulting from

repeated shifts of attention to locations would be dispersed among the various chunks representing the locations in which the object appeared, resulting in a decrease the probability of retrieving any of the multiple location chunks. Second, to the extent that interface objects are represented in a manner similar to the color-button chunks used in the model, (i.e., including a *loc* slot containing the location of the object), there would be a distinct object chunk for each location in which the object resided; because the contents of the *loc* slot would not be identical, the chunks would fail to merge. In turn, the existence of multiple memory representations of the object would not only disperse the base-level activation boost from repeated uses, but also could result in an erroneous shift of attention due the retrieval of an out-dated object, i.e., one containing an incorrect location chunk.

The performance advantages of using distinct and representative labels on interface objects were demonstrated in the data in the average button evaluation times shown in Figure 11. The advantage to having distinct labels becomes clear in a comparison between the performance of the no-label and arbitrary groups in this figure. In the model, the relative performance in the no-label condition suffers primarily due to the inability of this condition to rely on the faster label-matching strategy. To improve performance (i.e., not resort to accessing a ToolTip), this condition had to gradually acquire the ability to retrieve the chunks required by the location-recognition strategy. The difficulty in acquiring this ability was exacerbated by the large fan between the single, blank, chunk representing the (lack of a) button label, and the color-button chunks the model was trying to retrieve. Thus, distinct labels not only enable use of the

label-matching strategy, but also engender the advantages of having a unique retrieval cue.

The advantage to having object labels that are representative of the underlying function of the object is demonstrated in a comparison between the arbitrary and meaningful groups in Figure 11. Due to the close association between the colors and text labels, the model could always retrieve the target label in the meaningful condition and use it in conjunction with the display-based, label-matching strategy. In the arbitrary condition, the model could not use label-matching until it had learned the color to icon association; and learning this association occurred gradually, due to the initially weak relationship between the colors and icons. Thus, labels representative of the function of an interface object not only support early use of the label-matching strategy, but also prevent users from having to learn to associate weakly associated items.

The data and model presented in this research highlight the pervasiveness of location learning and the central role location knowledge plays in the skilled use of a graphical user interface. To the extent that interface object locations remain constant, users will eventually learn those locations and can use this location knowledge to limit visual search. If object locations vary, then location knowledge is rendered useless at best and misleading at worst, thus requiring a potentially time consuming exhaustive visual search of the screen with each episode of interaction.

Future Directions

Several interesting issues arose in this research that warrant further investigation. One major outstanding question concerns the nature of the approximate-

trials phenomenon found in Experiments I and II. The cause of the disproportionate number of tips, erroneous clicks and saccades to buttons adjacent to the correct ones was not unambiguously determined in the course of this research. Future research should attempt to tease apart the relative contributions of error in saccade accuracy versus error in the precision of location memory.

The presence of the approximate-trials phenomenon, and in particular the questions it raises about the precision of location memory, also points to a deeper theoretical issue related to the representation of locations in ACT-R/PM. At present, location chunks represent precise locations on the screen, using Cartesian coordinates. It may be the case that locations should be able to be represented in a less discrete manner, such as in zones, quadrants, or in qualitative categories such as “top” or “bottom”. It is quite likely that the ways in which people choose to spatially carve up the visual space is person- and task-specific (indeed there was some evidence for this in responses from participants in the Experiment I debriefing session). As such, a mechanism for functionally combining location chunks into higher order representations would be a means of maintaining the discrete, precise representation of locations, but also enabling the modeler to represent person- or task-specific higher order representations as appropriate. Determining the nature of these higher order representations could be an avenue of future research used to guide development of the vision module in ACT-R/PM.

REFERENCES

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

Anderson, J. R. (1991). Is human cognition adaptive? Behavioral and Brain Sciences(14), 471-517.

Anderson, J. R. (1993). Rules of the mind. Hillsdale, NJ: Erlbaum.

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

Anderson, J. R., Matessa, M., & Douglass, S. (1995). The ACT-R theory and visual attention, Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society (pp. 61-65): Hillsdale, NJ: Lawrence Erlbaum Associates.

Andrade, J., & Meudell, P. (1993). Short report: Is spatial information encoded automatically in memory? The Quarterly Journal of Experimental Psychology, 46A(2), 365-375.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. Journal of Cognitive Neuroscience, 7(1), 66-80.

Barreau, D., & Nardi, B. A. (1995). Finding and reminding: File organization from the desktop. SIGCHI Bulletin, 27(3), 39-43.

Blankenberger, S., & Hahn, K. (1991). Effects of icon design on human-computer interaction. International Journal of Man-Machine Studies, 35(3), 363-377.

Byrne, M. D., & Anderson, J. R. (1998). Perception and action. In J. R. Anderson & C. Lebiere (Eds.), The atomic components of thought (pp. 167-200). Mahwah, NJ: Lawrence Erlbaum.

Card, S. K. (1984). Visual search of computer command menus. In H. Bouma & D. G. Bouwhis (Eds.), Attention and Performance X: Control of Language Processes (pp. 97-108). Hillsdale, New Jersey: Erlbaum.

Card, S. K., Moran, T. P., & Newell, A. (1983). The psychology of human-computer interaction. Hillsdale, NJ: Erlbaum.

Coeeffe, C., & O'Regan, J. K. (1987). Reducing the influence of non-target stimuli on saccade accuracy: Predictability and latency effects. Vision Research, 27(2), 227-240.

Franzke, M. (1994). Exploration and experienced performance with display-based systems (#94-05). Boulder, Colorado: Institute of Cognitive Science, University of Colorado.

Franzke, M. (1995). Turning Research into Practice: Characteristics of Display-Based Interaction, Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems (Vol. 1, pp. 421-428).

Gray, W. D. (in press). The nature and processing of errors in interactive behavior. Cognitive Science.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. Journal of Experimental Psychology: General, 108(3), 356-388.

Howes, A. (1994). A model of the acquisition of menu knowledge by exploration, Proceedings of ACM CHI'94 Conference on Human Factors in Computing Systems (Vol. 1, pp. 445-451).

Howes, A., & Payne, S. J. (1990). Display-based competence: Towards user models for menu-driven interfaces. International Journal of Man-Machine Studies, 3, 637-655.

Howes, A., & Young, R. M. (1996). Learning consistent, interactive, and meaningful task-action mappings: A computational model. Cognitive Science, 20, 301-356.

Jones, W. P., & Dumais, S. T. (1986). The spatial metaphor for user interfaces: Experimental tests of reference by location versus name. ACM Transactions on Office Information Systems, 4(1), 42-63.

Kaptelinin, V. (1993). Item recognition in menu selection: the effect of practice, Proceedings of ACM INTERCHI'93 Conference on Human Factors in Computing Systems -- Adjunct Proceedings (pp. 183-184).

Kaptelinin, V. (1996). Creating computer-based work environments: An empirical study of Macintosh users. Proceedings of the 1996 ACM SIGCPR/SIGMIS conference, 360-366.

Kitajima, M., & Polson, P. G. (1997). A comprehension-based model of exploration. Human-Computer Interaction, 12(4), 345-389.

Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? Evidence from the tower of hanoi. Cognitive Psychology, 17, 248-294.

Lansdale, M. W. (1991). Remembering about documents: Memory for appearance, format, and location. Ergonomics, 34(8), 1161-1178.

Lebiere, C., & Anderson, J. R. (1998). Cognitive arithmetic, The atomic components of thought (pp. 297-342). Mahwah, New Jersey: Lawrence Erlbaum.

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

Lovelace, E. A., & Southall, S. D. (1983). Memory for words in prose and their locations on the page. Memory and Cognition, 11(5), 429-434.

Mayes, J. T., Draper, S. W., McGregor, A. M., & Oatley, K. (1988). Information flow in a user interface: the effect of experience and context on the recall of MacWrite screens. In D. M. Jones & R. Einder (Eds.), People and Computers IV . Cambridge, UK: Cambridge University Press.

Mehlenbacher, B., Duffy, T. M., & Palmer, J. (1989). Finding information on a menu: Linking menu organization to the user's goals. Human-Computer Interaction, 4(3), 231-251.

Morton, J. (1967). A singular lack of incidental learning. Nature, 215, 203-204.

Moyes, J. (1994). When users do and don't rely on icon shape, Proceedings of ACM CHI'94 Conference on Human Factors in Computing Systems (Vol. 2, pp. 283-284).

Moyes, J. (1995). Putting icons in context: the influence of contextual information on the usability of icons. Unpublished Doctoral Thesis, University of Glasgow, Glasgow.

Naveh-Benjamin, M. (1987). Coding of spatial location: An automatic process? Journal of Experimental Psychology: Learning, Memory and Cognition, 13(4), 595-605.

Naveh-Benjamin, M. (1988). Recognition memory of spatial location information: Another failure to support automaticity. Memory and Cognition, 16(5), 437-445.

Newell, A. (1990). Unified theories of cognition. Cambridge, MA, USA: Harvard University Press.

Nickerson, R. S., & Adams, M. J. (1979). Long-term memory for a common object. Cognitive Psychology, 11, 287-307.

Norman, D. A. (1988). The psychology of everyday things. New York: Basic Books.

O'Hara, K. P., & Payne, S. J. (1998). The effects of operator implementation cost on planfulness of problem solving and learning. Cognitive Psychology, 35, 71-98.

Payne, S. J. (1991). Display-based action at the user-interface. International Journal of Man-Machine Studies, 35, 275-289.

Payne, S. J., & Green, T. R. G. (1986). Task-action grammars: a model of the mental representation of task languages. Human-Computer Interaction, 2(2), 93-133.

Polson, P. G., & Lewis, C. H. (1990). Theory-based design for easily-learned interfaces. Human-Computer Interaction, 5, 191-220.

Postma, A., & DeHaan, E. H. F. (1996). What was where? Memory for object locations. The Quarterly Journal of Experimental Psychology, 49A(1), 178-199.

Rieman, J., Young, R. M., & Howes, A. (1996). A dual-space model of iteratively deepening exploratory learning. International Journal of Human-Computer Studies, 44(6), 743-775.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. Psychological Review, 84(1), 1-66.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. Psychological Review, 84(2), 127-190.

Simon, H. A. (1975). The functional equivalence of problem solving skills. Cognitive Psychology, 7, 268-288.

Smelcer, J. B., & Walker, N. (1993). Transfer of knowledge across computer command menus. International Journal of Human-Computer Interaction, 5(2), 147-165.

Somberg, B. J. (1987). A comparison of rule-based and positionally constant arrangements of computer menu items, Proceedings of ACM CHI+GI'87 Conference on Human Factors in Computing Systems and Graphics Interface (pp. 255-260).

Stratton, G. M. (1917). The mnemonic feat of the Shass Pollak. Psychological Review, 24, 244-247.

Suchman, L. A. (1987). Plans and situated action: The problem of human-machine communication. New York: Cambridge University Press.

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

Triesman, A., & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. Journal of Experimental Psychology: General, 114(3), 285-310.

Vandierendock, A., Hoe, R. V., & Soete, G. D. (1988). Menu search as a function of menu organization, categorization, and experience. Acta Psychologica, 69, 231-248.

Viviani, P., & Swensson, R. G. (1982). Saccadic eye movements to peripherally discriminated visual targets. Journal of Experimental Psychology: Human Perception & Performance, 8(1), 113-126.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. Cognitive Science, 18(87-122).