# Modeling How Menu Search Strategies Develop with Experience

Gilles Bailly[a,*], Daniel Duarte[b], Antti Oulasvirta[c], Luis A. Leiva[b]

[a] *Sorbonne Université, CNRS, ISIR, France*
[b] *University of Luxembourg, Luxembourg*
[c] *Aalto University, Finland*

**Abstract**

To find an item in a menu, users can follow different visual search strategies, such as scanning items one by one (serial search) or trying to remember where the item was (recall search). However, building predictive models of search behavior has turned out to be challenging, because these strategies evolve with practice. To address this challenge, we study theory-inspired models of visual search in linear menus and propose a novel arbitration mechanism to coordinate the adoption of such visual search strategies. Given a menu design and the user's previous experience with it, our approach predicts *when* different search strategies (serial, recall, random) will be adopted and *which* menu item will be fixated next. Our results (1) describe empirical data plausibly with psychologically valid and interpretable models, (2) provide new insights about how search strategies evolve with practice, and (3) show how to infer search strategy from eye tracking data. To sum up, the models provide a foundation to better understand how users learn to scan linear menus.

*Keywords:* Linear menus, Learning, Decision making, Visual search, Computational models

[*] Corresponding author

*Email addresses:* gilles.bailly@sorbonne-universite.fr (Gilles Bailly), daniel.tojal@live.fr (Daniel Duarte), antti.oulasvirta@aalto.fi (Antti Oulasvirta), luis.leiva@uni.lu (Luis A. Leiva)

## 1. Introduction

Computational models of user performance have applications in the design, adaptation, and personalization of graphical user interfaces (GUIs) [6, 12, 16, 19, 23, 47, 58, 63]. Models that are developed from first principles—that is, when they are informed by theory—can efficiently synthesize complex phenomena in an explainable form. This affords scrutiny by practitioners and supports the development of theoretical ideas. However, modeling remains a challenging task, mostly because interacting with GUIs involves multiple and subtle phenomena involving contributions from motor control, learning, and decision making.

In this paper, we study computational models of visual search in linear menus. We focus on menus because they are among the most pervasive interfaces for selecting options in GUIs. They can be found in a wide range of devices (e.g., desktop, mobile, or web interfaces) and tasks (e.g., selecting a command, an application, or a service) [4].

In fact, there is a broad body of research to understand and improve menu interaction, including theories, empirical studies, models, and interaction techniques [4]. For example, several dozens of interaction techniques have been proposed [3], including model-based techniques such as adaptive menus [12, 19, 58, 63] or menu optimisation [6, 16, 23, 47]. All of these applications would benefit from more advanced models of visual search. Although numerous studies have investigated visual search in menus [5, 7, 48, 50, 59], existing models do not precisely synthesize the corresponding phenomena, mostly because they are idiosyncratic (i.e., specific to each individual) and depend on a number of factors such as the design of the menu.

We study seven computational models to better understand visual search in menus, described in Section 3. The models simulate how visual search strategies (random, serial, and recall search) evolve with users' practice. Realizing that users may strategically change their strategy, we propose a hierarchical arbitration mechanism. It is a higher-level principle that coordinates the selection of a visual search strategy given experiences so far. This allows characterizing user's skill with a given menu in a new way, accounting for how they transition between search strategies. A key aspect of our work is that the studied models are theory-inspired and human-interpretable, unlike black-box machine learning models, such as neural nets.

Our modeling approach is inspired by decision-making theory from neuroscience and empirical findings of menu performance from HCI. Critically,

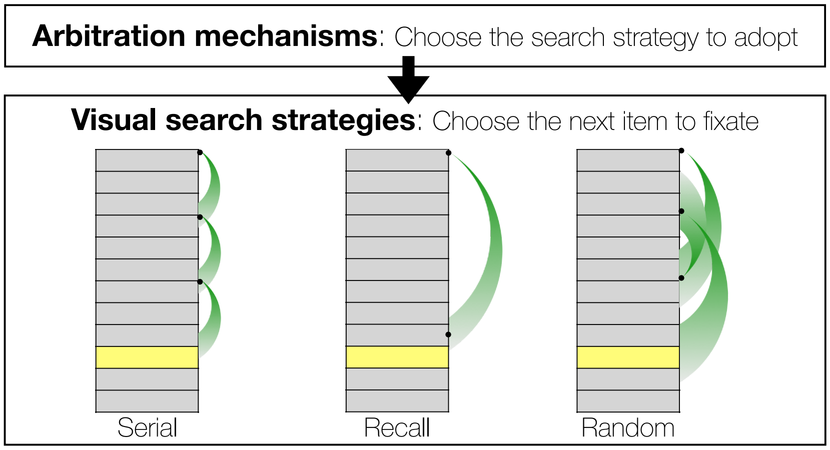

**Arbitration mechanisms**: Choose the search strategy to adopt
**Visual search strategies**: Choose the next item to fixate
Serial
Recall
Random

Figure 1: Overview of models. We consider three lower-level visual search strategies: Serial, Recall and Random search. These consist of choosing the item to fixate next given a visual search strategy. However, also the choice of these strategies can evolve with user's practice, represented here by an arbitration mechanism choosing a visual search strategy.

our studied models not only predict but also *explain* how users become more skilled over time. More precisely, our computational problem is formulated as follows: The input to all models are (i) the menu itself (how items are arranged), (ii) a target menu item, and (iii) the user's previous exposure to the menu, if any. The task is to output the probability of adopting one of the three search strategies (random, serial, and recall) and derive the sequence of most plausible fixated items (also called *menu scanpath*). To the best of our knowledge, our work is the first in tackling this challenging modeling perspective, i.e. to understand the extent to which menu scanpaths can be predicted and explained by hierarchical decision making. Our contribution is threefold:

1. *Theoretical foundations.* We present seven computational models of visual search in menus. A key aspect of our approach is that we take inspiration from human goal selection as a hierarchical control problem [49], where a higher level controller has several lower level controllers to choose from, as illustrated in Figure 1. At the *high level*, we propose an *arbitration mechanism* that chooses which one of the three search strategies (random, serial, and recall) a user is likely to adopt in order to locate a target item. The probability of each strategy evolves with practice. At the *low level*, the models predict a probability

distribution, which for each item, tells us the probability it will be the next one to be fixated. The models can thus derive the full sequence of fixated items until finding the target item, and ultimately to derive search time or other second-order measures.

2. *Methodological insights.* A key challenge in the study of search strategies in menus is being able to infer them from a sequence of observations on where the user has fixated. While the sequences of fixated items (sequences of x,y coordinates) are directly accessible from an eye-tracker, the adopted strategies (e.g., random, serial, or recall) are difficult to reconstruct either computationally, visually (by human labeling), or even verbally (by asking the participants) as illustrated in Figure 2. One of the proposed models we studied (Maximal) can infer automatically the most plausible search strategy from a sequence of fixated items.
3. *Empirical evaluation.* We compare the seven models on a comprehensive dataset [5] that contains sequences of fixated items, captured by a high-quality eye tracker, for different menu organizations (Unordered, Alphabetic, and Semantic) and different menu lengths (8, 12, and 16 items). We rely on one of the model (namely Maximal model) to provide new insights about the data and the visual search strategies. Our results suggest that (i) the three search strategies are necessary to explain users' behavior and that (ii) they evolve with practice. In particular, the proportion of recall search increases linearly with practice but is not as prominent as expected. Menu organization also influences the evolution of visual search strategies. For example, the decrease of serial search is more prominent for the Unordered menu organization.

## 2. Related Work

Menu interaction in general and visual search in linear menus in particular have gained significant attention in the field of HCI [3, 20, 26]. We note that, while our work focuses on menus and visual search, it may be of interest to a broader question of how users develop skills when interacting with other types of GUIs such as forms.

### *2.1. Visual search in menus*

Designing menus is a frequent yet challenging task, especially when the number of commands increases, because designers need to address multiple

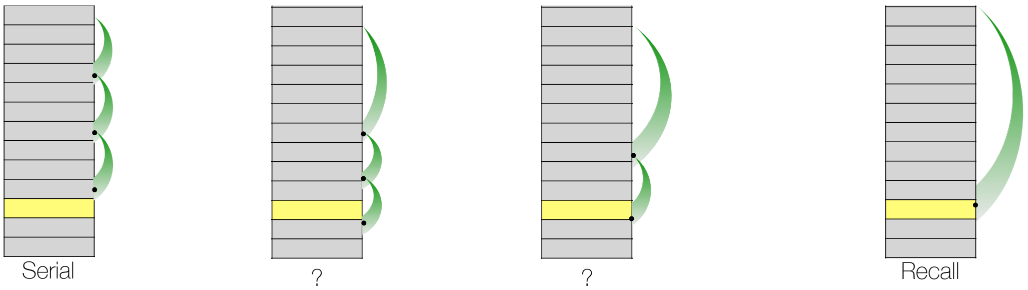


Figure 2: Four sequences of item fixations. While it is easy to label the leftmost (Serial) and rightmost (Recall) visual search strategies, the two central ones are much more difficult. They could be a Serial search (systematic top-down inspection) with some variability about the distance between two consecutive fixations or a Recall search with some uncertainty about the exact target location. Our method relies on our model (Maximal) and a training set to infer the most likely visual search strategy from a sequence of fixated items.

objectives such as speed, accuracy, satisfaction, learnability, etc. [6]. While the common approach relies on trial and error, or heuristics, several systems have been proposed to automatically evaluate the quality of a menu system and optimize their design [6, 16, 23, 47]. Adaptive menus also have the potential to improve usability [58]. They autonomously change their content [42], organization [58, 62], or style (e.g., color, size, etc.) [12, 19, 63] by considering users' capabilities and interests. In either case, user models are required to predict the impact of a change on user behavior. In particular, models of visual search strategies are especially important for adaptive menus in order to place/move relevant items at locations where the user is likely to look next. However, elaborating robust mathematical models of human behavior, beyond simple tasks like pointing (Fitts' law [21]) remains challenging [16].

### *2.1.1. Empirical studies*

Numerous empirical studies have investigated the influence of various factors on user behavior with menus [3]. These factors include: menu length (total number of items), item location, menu organization, interactive mechanisms, and/or user expertise [5, 7, 48, 59]. These studies generally show subtle interaction effects on search time, but only few of them have investigated their impact on fixated items or gaze distribution [5, 7]. For example, Byrne et al. [7] shows how target location influences the gaze distribution. However, the data is aggregated over trials and individuals. It is thus difficult to interpret which combination of visual search strategies lead to these gaze distributions.

To the best of our knowledge, none of these studies report which visual search strategy is adopted and how these factors influence visual search strategies, probably due to the lack of methods to reconstruct these strategies from a sequence of fixated items.

### *2.1.2. Computational models*

Various models of user performance with menus have been proposed [11, 17, 23, 44, 45] but few of them made visual search strategies explicit. Early models considered two visual search strategies: **serial** search, i.e., a top-down exploration [7, 29, 41], and **random** search [7, 28]. These models, however, did not comprehensively describe the observed data, mostly because they did not take into account the effect of user expertise on visual search.

Cockburn et al. [12] took into account the effect of user expertise by assuming a logarithmic transition from visual-based search (novice users) to decision-based search (expert users). The precise nature of the visual search itself was not specified, but it is likely to be random search as they reported that *"visual search time is linear with the total number of items"* (but not the location of the target item). They also found that decision time was linear with the entropy of each item.

Later, Bailly et al. [5] considered that visual search is a combination of serial search and **directed** search. Directed search shares some similarities with the work of Cockburn et al. [12]; it is initially a random search [53] but, with practice, users progressively restrict the visual search to the area around the target. When the user is an expert, the directed search becomes **recall** search, where users directly fixate on the target item. The learning component relies on the Power Law of Practice (PLP) [51].

Recently, Todi et al. [62] considered three visual search strategies: serial, recall, and **semantic** search. However, their approach does not explain how these visual search strategies evolve over time. Chen et al. [8] used a different approach rooted on the computational rationality theory, where they did not make assumptions regarding search strategies. Instead, they are supposed to emerge when an eye-movement policy is optimized to consider task rewards, low-level perceptual constraints (e.g., saccade duration), and design constraints (e.g., menu organization). However, the model assumes a single search strategy per menu organization that does not evolve with user expertise.

### *2.2. Scanpath models*

Our work is also related to scanpath models that predict a sequence of fixations. In this regard, we should mention that several models have been proposed [37] and tested in different contexts such as natural scenes [13], virtual reality [30], graphical layouts [33], hierarchical structures [60] or information visualizations [65]. Our work differs from this line of research in several ways. First, we focus on menu interaction, which has its own characteristics, namely: well structured information without particularly salient elements that are however frequently revisited. Second, we study not only the sequence of fixations (low-level, observable) but also the cognitive process that generates the observable data; that is, the search strategies (high-level, non-observable) that produce such a sequence of fixations. Also, and more important, we study how users' behavior evolve over time.

Scanpath models are evaluated by comparing the generated fixation sequences against the human ones, using some (dis)similarity metric such as DTW, Eyeanalysis, Multimatch, etc. However, this approach aggregates data and tends to ignore the underlying generative cognitive processes [32, 37]. A more nuanced approach, which is state-of-the-art in cognitive science, consists of treating each fixation in the scanpath as a decision and evaluating how well the model predicts this decision given the previous ones using the log-likelihood metric [57]. We also use it to evaluate our models (see Section 4.3 for more details) as it has been proved successful in many other works [10, 38].

### *2.3. Learning and decision-making*

Finally, our work is related to decision-making and learning models, since we focus on modeling how users choose a search strategy and the item to fixate next (decision) and how these decisions evolve with practice (learning). Surprisingly, previous work has seldom consider these two phenomena together. Bailly et al. [2] presented a model describing the learning dynamics of keyboard shortcut adoption. An agent chooses among using the *menu* to execute the command, using the *shortcut* to execute the command or *learning* the mapping between the command and its corresponding shortcut. The choice of these high-level strategies depends on several cognitive mechanisms including implicit/explicit memory or the ability of the users to consider several actions ahead (planning). Li et al. [46] presented a computational model of menu navigation for blind people. The agent chooses among three navigation actions—swiping, gliding and direct touch—depending on user familiarity with the actions, the menu organization, their memory, etc. Our work is inspired by these models by formulating visual search in menus as a learning and decision-making problem, but focuses on different types of decisions (search strategy and fixated items instead of interaction techniques).

### *2.4. Summary*

Previous work has proposed several models of visual search, for example for menu optimisation and adaptation, however there is no consensus on how users adopt visual search strategies and how they evolve over time. One reason is that previous work has been directed at modeling and/or predicting performance (time) and/or gaze distribution on menus, ignoring the sequence of fixated items (menu scanpath), which is actually the only way to precisely study visual search strategies [57]. The same time (or gaze) distribution can result from different combinations of visual search strategies. Moreover, collected data in previous work has been generally aggregated over trials and/or participants, thereby masking individual decision-making processes regarding important questions such as "which visual search strategy to adopt" or "which item to fixate next". For example, Bailly et al. [5] collected an extensive dataset providing the sequence of fixated items for each trial in several menu configurations, however they did not indicate which visual search strategy was used by the participants. In sum, this paper advances our collective understanding of visual search in linear menus by (1) quantifying the adoption of three main visual search strategies (random, serial, recall) and studying their evolution over time; (2) making predictions about the visual search strategies and deriving the sequences of fixated items for each each individual; and (3) comparing these predictions against empirical data.

## 3. Modeling approach

A key aspect of our modeling approach is that we look at *learning*: we aim to explain and predict how visual search strategies, as informed by their corresponding sequences of eye fixations, evolve over time. Therefore, based on previous work [5, 12], we assume that users should gradually become experts as long as they keep interacting with a given type of menu (e.g., where items are always sorted alphabetically) over several trials. Typically, novice users would adopt a serial or random search because they do not know where the target items are [5]. In contrast, more experienced users remember where each target item is and therefore should adopt a recall search strategy [12, 62]. It remains that users do not always become experts because it might be easier to rely on recognition than memory [2, 24, 39, 67]

### *3.1. Problem formulation*

We frame visual search as a hierarchical 2-level discrete-time stochastic control for decision making. At the **high-level** decision-making process, users rely on an arbitration mechanism (an *arbitrator*, for short) to choose a visual search strategy. We currently consider three possible visual search strategies, based on previous work [7, 29, 41, 62]: serial search, recall search, and random search. We assume that the arbitration mechanism depends on user expertise.

Once the visual search strategy is (unconsciously) chosen, users have to decide which item to fixate next. This is a **low-level** decision-making process driven by the selected strategy. We assume that the next item to fixate on does *not* depend on the user expertise, but rather on the chosen strategy and the currently fixated item. The assumption that users do not change strategy *during* a visual search is acceptable given the fact that we consider scenarios where the target item is well defined (e.g., a text label) and the menus are relatively small (e.g., 16 items at most).

We first describe the implementation of the three visual search strategies (low-level). This is important in order to properly contextualize the whole visual search process. We then describe and compare different arbitration mechanisms (high-level) to choose a visual search strategy. The key notations and parameters are summarized in Tables 1 and 2.

### *3.2. Low-level decision making: Which item to fixate next?*

Let $n$ denote the number of items in the menu, or its *length*. The choice of the item to fixate next depends on the target menu item $T \in \{1, n\}$ and the currently fixated item $F_t \in \{1, n\}$, i.e., the item the user is currently looking at in the menu. We then define $s_t \in \mathcal{S}$ be a *state* defined by a tuple $(T, F_t)$ at time $t$. The number of all possible states depends on the menu length $n$.

Given a state $s_t$, the user chooses an action $a_i \in \mathcal{A}$. In our context, the set of actions are the items the user can fixate on, plus a terminal action which indicates that the visual search process has finished. In a menu with $n$ items, there are thus $n + 1$ actions. We use $a_i$ to indicate the action of fixating on item $i \in \{1, n\}$ and $a_e$ to indicate the terminal action, where the user either finds the target item ($e = T$) or leaves the application ($e \neq T$).

The terminal action is automatically chosen when the user is fixating on the target item.

The input of this low-level decision making process is a state $s_t$ and its output is the chosen action (either $a_i$ or $a_e$). The model iteratively predicts the sequence of fixated items for a given visual search strategy. We now describe the three visual search strategies considered.

### *3.2.1. Serial search*

A serial search strategy is a systematic top-down inspection of the menu, or bottom-up inspection when reaching the end of the menu, until the target item is found. Given a currently fixated item $F_t$, the next fixated item $F_{t+1}$ is estimated as:

$$F_{t+1} = F_t + A_{\text{serial}} + \mathcal{N}\left(0,\, \sigma_{\text{serial}}\right) \tag{1}$$

indicating that the saccadic distance between the current fixation point and the landing fixation point is $A_{\text{serial}}$ with a certain amount of uncertainty represented by a Gaussian distribution with a standard deviation $\sigma_{\text{serial}}$. We expect $A_{\text{serial}}$ to be the number of items covered by the fovea. We then derive the probability to fixate on item $i$ given the current state $s_t = (T, F_t)$:

$$P(a_i|s_t) = \frac{\mathcal{N}\left(a_i, F_t + A_{\text{serial}},\, \sigma_{\text{serial}}\right)}{\sum_k^n \mathcal{N}\left(a_k, F_t + A_{\text{serial}},\, \sigma_{\text{serial}}\right)} \tag{2}$$

This model has two parameters: $A_{\text{serial}}$ and $\sigma_{\text{serial}}$.

### *3.2.2. Recall search*

A recall search strategy consists of directly fixating the target item without having to inspect the entire menu [62]. Similar to the serial search strategy, there is some variability characterizing the uncertainty around the exact location of the landing fixation point. This variability is represented by a Gaussian distribution $\mathcal{N}\left(B_{\text{recall}},\, \sigma_{\text{recall}}\right)$ where $B_{\text{recall}}$ is a systematic bias. We expect $B_{\text{recall}}$ to be positive to reflect that saccadic movements tend to undershoot the visual target [27] and small enough so that the target item remains in the field of vision centered on the landing fixation point.

$$F_{t+1} = T + \mathcal{N}\left(B_{\text{recall}},\, \sigma_{\text{recall}}\right) \tag{3}$$

We then derive the probability to fixate the item $i$ given the current state $s_t = (T, F_t)$:

$$P(a_i|s_t) = \frac{\mathcal{N}\left(a_i, T - B_{\text{recall}},\, \sigma_{\text{recall}}\right)}{\sum_k^n \mathcal{N}\left(a_k, T - B_{recall},\, \sigma_{\text{recall}}\right)} \tag{4}$$

This model also has two parameters: $B_{\text{recall}}$ and $\sigma_{\text{recall}}$.

### *3.2.3. Random search*

The literature on menu interaction [53] generally distinguishes between:

- random search without replacement: a given item can only be visited once.
- random search with replacement: a given item can be visited several times.

Our random search model includes a *degree of replacement* to generalize the two previous strategies. Moreover, it modulates the degree of replacement for a given item depending on when it was last visited. It is inspired by the theory of inhibition of return [36]: users are less likely to revisit an item if it has recently been visited.

From the history of previous eye fixations, we get $\mathcal{V} = \{v_1, ...v_n\}$ a vector of size $n$ representing for each item $i$ in the menu, the number of fixated items from its last visit. $v_i$ is bounded between 0 and $V_{\max}$ where $V_{\max}$ is a parameter reflecting the maximum number of fixations that users can remember. We use the Boltzman soft-max function to convert the $v_k$ values into a probabilistic action:

$$P(a_i|\mathcal{V}) = \frac{e^{\beta v_i}}{\sum_k e^{\beta v_k}} \tag{5}$$

This model has two parameters: $V_{\max}$ and $\beta$. The $\beta$ parameter is the inverse temperature which controls the degree of replacement; i.e., a small value of $\beta$ reflects almost "pure" random choice (i.e, with replacement), where each item has the same probability to be fixated regardless previous fixations. A high value of $\beta$ reflects that an item which has been fixated is unlikely to be revisited, i.e., a random search without replacement.

### *3.3. High-level decision-making: Which visual search strategy to select?*

In the previous section, we described how the next fixated item is chosen given a visual search strategy (low-level). We now present a series of arbitration mechanisms (high-level) to decide *which* visual search strategy to adopt to reach a given target. The input of an arbitration mechanism is a command (e.g., 'Open File') denoted as $s'_t \in \mathcal{S}'$ and the output is the chosen strategy

Table 1: Key notations for the low-level decision making models.

| Notation | Description |
|---|---|
| $n$ | number of items in the menu (menu length) |
| $i$ | $i^{th}$ menu item, $i\ in\{1,n\}$ |
| $T$ | Target menu item: item users have to select |
| $t$ | Time-based index of fixation |
| $F_t$ | Currently fixated item ,ie., the item the user is currently looking at in the menu at time $t$ |
| $s_t = (T, F_t)$ | State indicating the target item $T$ and the current fixation $F_t$ |
| $a_i$ | The action of fixating on item $i$ |
| $\mathcal{V}$ | $\mathcal{V} = \{v_1, ..., v_n\}$ a vector size $n$ representing for each item $i$ in the menu, the number of fixated items from its last visit. |

$a' \in \mathcal{A}'$ for this command. The three considered visual searches $a'$ are serial, recall and random search. By iterating on the sequence of commands[1] to execute, we can then derive the sequence of visual search strategies and study how they evolve over time.

We introduce four classes of arbitration mechanisms, resulting in a total of 7 models:

- **X-only**, where X $\in$ {Serial, Recall, Random}. These arbitration mechanisms systematically select one search strategy; e.g., Serial-only always selects Serial search. These three X-only arbitration mechanisms do not really "arbitrate" as they systematically select the same strategy (the serial arbitration mechanism always selects the serial search strategy).

- **Random**. This arbitration mechanism randomly selects a visual search strategy among all those available (serial, recall, random); i.e., all strategies have the same probability to be chosen by the user. "Random-only" and "Random"should not be confused. The former systematically selects the (low-level) random visual search strategy. The latter is a high-level arbitration mechanism which randomly selects one strategy among all available on each trial, including the low-level random

[1] The probability of transitioning to another state $s'_{t+1}$ depends on the frequency of the command usage; i.e., it does not depend on the user's behavior.

Table 2: Free parameters of the three low-level models of visual search strategy. The range of the free parameters is the one used to fit the models.

| **Symbol** | **Range** | Model | **Description** |
|---|---|---|---|
| $A_{\text{serial}}$ | $[0, n]$ | Serial | Average distance performed by a saccade in a systematic top-down inspection of the menu |
| $\sigma_{\text{serial}}$ | $[0, 100]$ | Serial | Standard deviation. The uncertainty about the landing point in the serial search strategy is represented by a Gaussian distribution with a standard deviation $\sigma_{\text{serial}}$. |
| $B_{\text{recall}}$ | $[0, n]$ | Recall | Systematic bias. The uncertainty around the exact location of the landing fixation point in the recall search strategy is represented by a Gaussian distribution centered on $B_{recall}$. |
| $\sigma_{\text{recall}}$ | $[0, 100]$ | Recall | Standard deviation in the recall search strategy. The uncertainty around the exact location of the landing fixation point is represented by a Gaussian distribution centered on $B_{recall}$. |
| $V_{\max}$ | $[0, 10]$ | Random | Maximum number of fixations (visited items) a user can remember within a single trial. |
| $\beta$ | $[0, 10]$ | Random | Inverse temperature which controls the degree of replacement; i.e., a small value of $\beta$ reflects almost "pure" random choice (i.e, with replacement), where each item has the same probability to be fixated regardless previous fixations. A high value of $\beta$ reflects that an item which has been fixated is unlikely to be revisited, i.e., a random search without replacement. |

search strategy.

- **Fluid** and **Memory**. These arbitration mechanisms aim to explain the novice-to-expert transition; i.e., a user who starts using Serial search and gradually adopts Recall search over time.
- **Maximal (upper bound)**. This mechanism assumes that the user always adopts the best strategy, considering the observed data.

The three **X-only** arbitration mechanisms as well as **Random** primary serve as baselines for model comparison. They are unlikely to well reflect general behaviors as it is known that users tend to transition from novice to expert behavior; i.e, they tend to adopt Recall search with enough practice [5, 11]. They thus provide bottom-line performance estimates (in particular, the Random arbitration mechanism provides a chance-level estimate). We consider a model successfully predicts/explains users behavior if it is substantially better (in terms of log-likelihood and BIC) than these baseline models. A key feature of our evaluation (Section 4) is to distinguish population-level and individual-level models, which is important to address user variability in decision-making problems [34]. For example, it is possible that some users only use serial search strategy (Serial-only arbitration) due to a lack of motivation, cognitive limitation, or simply because the experiment was either too short or too complex to memorize item locations. For example, Grossman et al. [24] noted that 50% of the participants does not adopt expert behavior.

**Fluid** and **Memory** are introduced to characterize more precisely the transition from novice to expert behavior. One simplification of both models is to not cover random search, i.e., considering the transition from serial to recall search. The reason for this simplification is that the proportion of random search is (1) less prominent than the two other strategies and (2) depends less on user expertise. We discuss this modeling assumption in more detail in Section 6.

Finally, the **Maximal** arbitration mechanism picks the strategy that is the closest match with the observed behavior. For example, if the user seems to jump directly close to the target, it is most likely to be an instance of recall search, while if there are lots of jumps, it may be closer to random search. More precisely, maximal arbitration picks the strategy that best explains the observed data for a given command $s'_t$ and a given participant. The main difference from the previous mechanisms is that this one cannot be fully simulated, since this mechanism requires knowledge of *all* the interactions performed by the user. However, it has three major advantages:

- it provides an upper bound for model comparison, while the previous ones provide bottom-line performance estimates.
- it reconstructs a visual search strategy from a sequence of fixated items. As illustrated in Figure 2, human visual inspection is not sufficient to accurately label a search strategy from a sequence of fixated items.
- the estimation of the parameters of this model (Section 4.3.2) is more psychologically plausible than the other models, by construction: an upper-bound model considers the strategy which best fits the empirical data for each command selection.

### *3.3.1. Implementation*

**Serial-only** has two parameters, those of the serial visual search: $A_{\text{serial}}$ and $\sigma_{\text{serial}}$, but they have the same value for all trials of a given user, i.e., they are *intrinsic* to every user.

**Recall-only** has the parameters of the recall visual search: $B_{\text{recall}}$ and $\sigma_{\text{recall}}$, also intrinsic to every user.

**Random-only** has the parameters of the random visual search: $V_{\text{max}}$ and $\beta$, also intrinsic to every user.

The **Random** arbitration mechanism randomly selects a strategy among the three visual search strategies:

$$P(\text{recall}) = P(\text{serial}) = P(\text{random}) = 0.33 \tag{6}$$

It has six parameters, two per low-level model involved: $A_{\text{serial}}$, $\sigma_{\text{serial}}$ , $B_{\text{recall}}$, $\sigma_{\text{recall}}$, $V_{\text{max}}$, $\beta$.

The **Fluid** arbitration mechanism aims to explain the novice-to-expert transition, i.e., a user who switches progressively from serial to recall search strategy. The model follows a sigmoid function:

$$P(\text{recall}|s'_t) = \left(1 + e^{\frac{-(t-a)}{b}}\right)^{-1} \tag{7}$$

where $a$ and $b$ are two parameters reflecting how quickly users switch from serial to recall, and $t$ the current trial. Conversely:

$$P(\text{serial}|s'_t) = 1 - P(\text{recall}|s'_t) \tag{8}$$

We chose a sigmoid function because it offers more flexibility than the logarithmic function. Typically, when $a$ is close to 0, the two functions are very similar, but the sigmoid function better captures the behavior of users who started the novice-to-expert transition late.

The **Memory** arbitration mechanism also aims to explain the novice-to-expert transition, but here the probability of choosing the recall strategy for a given target item $T$ depends on how strongly this item is stored in the user's memory. To estimate this strength, we use the base-level activation from ACT-R [1]:

$$B(T) = \log \sum_{j=1}^{k} (t - t_j)^{-\rho} \tag{9}$$

where $B(T)$ is the level of activation of target item $T$ in the memory, $t$ is the current trial, $t_j$ are the trials where $T$ has been selected and $k$ is the number of times $T$ has been selected. Finally, $\rho$ is a decay parameter that allows forgetting over time. The probability of choosing recall search is :

$$P(\text{recall}|s_t) = \left(1 + e^{\frac{-(B(T)-a)}{b}}\right)^{-1} \tag{10}$$

Similarly, the probability of choosing serial search is given by Equation 8.

Finally, the **Maximal** arbitration mechanism selects the strategy that best explains the observed data for each trial $t$ of a participant $p$. That is, it will choose $a_t^*$ such as:

$$a_t^*(p) = \arg\min_{a^*}(\mathcal{L}_{a^*}) \tag{11}$$

where $\mathcal{L}_{a^*}$ is the fitness function that reflects how well a strategy (low-level model) predicts the sequence of fixations performed by participant $p$ at trial $t$. The implementation of the fitness function $\mathcal{L}_{a^*}$ is described later in Section 4.3.2. Like in the Random arbitration mechanism, the Maximal arbitrator has six parameters per user, two per visual search strategy: $A_{\text{serial}}$, $\sigma_{\text{serial}}$ , $B_{\text{recall}}$, $\sigma_{\text{recall}}$, $V_{\text{max}}$, $\beta$.

## 4. Evaluation

We evaluated the models against an often used dataset on menu selection performance [5]. We chose this dataset as it replicates and extends previous datasets of menu interaction, e.g. [52]. Moreover, it is perhaps the most extensive open datasets, in terms of independent variables (menu length, menu organization, target item, practice) and dependent variables (time, accuracy, fixated items and gaze points). Importantly, it allows studying learning effect by providing enough repetition for each condition.

### *4.1. Experiment design*

The experimental design is fully described in [5]. Here we summarize the main characteristics of the study, that are relevant to our work, and the collected data. Twenty-two participants took part in a study about target selection on menus. However, due to missing data about gaze fixations for some users and conditions, 19 participants were eventually retained for analysis. The user interface prompted participants with a simple button with the name of the target item that should be found in different types of menus, as is common for these types of experiments. When participants clicked on that button, a menu appeared and they had to select the target item as fast as possible. In the case of missing items (i.e., the target item is not in the menu), the trial finished upon pressing the space bar. In our analysis, we do not consider the trials where the target item was not in the menu.

A 1750 Tobii eye-tracker was used (sampling rate: 50 Hz; latency: 20 ms; spatial resolution: 0.25 deg). The distance between the users' eyes and the screen was 65 cm. The height of each menu item was 0.75 cm and menu separators were 0.1 cm each. It means that participants could simultaneously focus on about 3 items at once, assuming a human fovea of 2 degrees [56]. The width of the menu is proportional to the largest word length in the menu.

The experiment compared 3 *stable* menu organizations (Unordered, Alphabetical, Semantic), 3 menu lengths (8, 12, 16 items) and target positions. These factors are the most studied ones in menu interaction as they are involved in multiple phenomena [5]. The items in the menu have less than 3 letters length difference to reduce potential saliency effects. Participants carried out at least 12 blocks per menu type, i.e., a given menu organization with a specific menu length. The menu content is repeated from trial to trial and from block to block to study does not change. Moreover, each item is selected the same number of times per block to study the impact of target location and practice on users' behavior.

The experiment followed a within-participants design: each participant tested the 9 conditions (3 organizations × 3 lengths) while findings all target items. The order of the conditions was counterbalanced between participants using a Latin square design. Figure 3 provides a diagram of the experiment workflow.

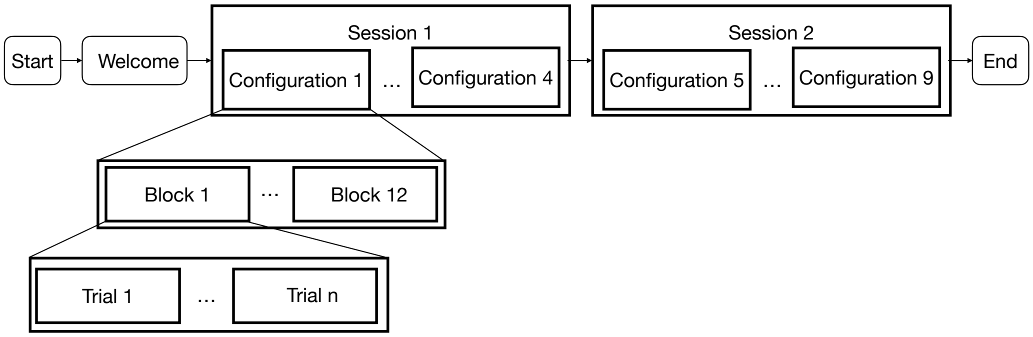


Figure 3: Workflow of the experimental design. Configuration is a menu length (8,12, 16) × a menu organisation (alphabetic, semantic, unordered). The presentation order of the configuration is counterbalanced between participants.

In total, the experiment comprised 19 participants × 3 menu organizations × 12 blocks × $\{8, 12, 16\}$ targets (depending on menu length) = 24,624 trials or command selections. The duration of the experiment was about 4h30m (3 sessions of 90 minutes each) per participant. It means about 4.8 selections per minute, considering that the average duration to select an item is about 1.5 s. The time between two sessions was from 0 to 7 days depending on the availability of the participants. Moreover, during the experiment, participants took a break every 50 selections. Breaks between conditions were also used to re-calibrate the eye-tracker. So, potential risks of fatigue, declining attention, or incorrect calibration were actually very limited.

### *4.2. Collected data*

In addition to completion times (time to find a target) and errors (whether the target was found or not), the dataset logs include the raw gaze paths as well as the sequence of gaze fixations and fixated items for each trial (Figure 8 shows some examples). In this paper, we focus on visual search and thus we only use the sequence of fixated items for analysis.

### *4.3. Evaluation methods*

In the following we describe the methods used to evaluate and compare the models (arbitration + low-level mechanisms).

#### *4.3.1. Metrics*

As discussed previously, we consider a menu scanpath as a *sequential* decision-making problem reflecting cognitive processes: How the history of past fixations influences the next user decision? In other words, we are more interested in the relative distribution of fixations rather than their absolute locations. For these reasons, we chose the log-likelihood metric [37, 57], which has the following characteristics.

- It is commonly used to evaluate dynamic models of cognition [57]. It relies on information theory, allowing not only to efficiently compare models (the full information in the model predictions is used for evaluation [37]) but also to estimate model parameters.

- It is especially appropriate to measure the plausibility of a scanpath among all possible ones despite the combinatorics (the number of sequences exponentially increases with sequence length) [57]. In our context, it measures the plausibility of a sequence of temporal ordered fixated menu items. A fixated item being the item located at the fixation point. However, our metrics does not take into account the absolute duration of fixation, which is acceptable as we focus on the decision-making problem.

- It is also appropriate to consider personalized, individual models (i.e., a different set of parameters for each participant) rather than a population model (the same set of parameters for all participants), which is important to address inter-individual variability in decision-making problems [34]. Different users can have radically different policies leading to different sequence of fixated items. Consider an extreme case with two users, one using Serial-only strategy and one using the Recall-only strategy: The notion of "average" user does not mean that they will use serial and recall strategies half of the time.

- It is part of the Bayesian framework, which is appropriate given the probabilistic nature of scan paths and the output of our model. Moreover, the Bayesian framework reconciles discrepancies between many scanpath metrics and does not require hyperparameters to be learned (common to classic scanpath models in eye-tracking research) [18, 37].

- Finally, by considering each fixation individually, the log-likelihood metric allows for detailed analyses to understand for example which fixations of the sequence cause problems in the modeling [37].

### *4.3.2. Log-likelihood implementation*

Our metric (or, more precisely, our fitness function) $\mathcal{L}$ reflects the capacity of each model to replicate participants' behavior [15, 57, 64]. In Bayesian terms, it is the likelihood of the data given the model, that is, the probability that the model chooses the same series of actions (here, the whole sequence of fixated items) as a given participant $p$. In practice, it consists of evaluating the likelihood of each participant's action $a_f^p$ given the past data $\{a_1^p, ..., a_{f-1}^p\}$. Here, it is important to specify that the past data includes all actions (fixated items) made by a participant $p$ from the beginning of the experiment,[2] not the actions made by the model.

$\mathcal{L}$ can be expressed as:

$$\mathcal{L}(m, p) = \prod_f P(a_f^p | \{a_1^p, ...a_{f-1}^p\}, m, p) \tag{12}$$

where $m$, $p$, and $f$ are respectively the model, the participant, and the index of the whole sequence of fixated items. A more detailed mathematical formulation of $\mathcal{L}(m, p)$ is available in Appendix A. We consider the **log-likelihood** $\log \mathcal{L}(m, p)$, which is equivalent to the likelihood computation but numerically more tractable [66], especially for very small numbers.[3]

### *4.3.3. Parameter estimation*

Parameter estimation consists of finding the set of model parameters $\hat{\theta}_m^p$ that best describes behavioral data, i.e., the set of parameters which maximises the log-likelihood:

$$\hat{\theta}_m^p = \underset{\theta}{\operatorname{argmin}} \log \mathcal{L}(m, p; \theta) \tag{13}$$

We use the differential evolution algorithm [54] as optimization method to find $\hat{\theta}_m^p$ for each model $m$ and each participant $p$.

### *4.3.4. Model comparison*

As previously stated, to compare models, we compute their log-likelihood with the best set of parameters $\hat{\theta}_m^p$. The model with the largest log-likelihood

[2] The sequence of fixated items of each trial are concatenated.

[3] The likelihood is a product of probabilities, whereas the log-likelihood equivalent is a sum of the log of such probabilities.

value is likely to better explain the observed data. However, it is advised to include a penalty term to account for model complexity, i.e., the more complex the model (in terms of number of parameters) the more it will tend to overfit [66]. For this, the Bayesian Information Criterion (BIC) is commonly used [9]:

$$\mathrm{BIC} = -2 \log \mathcal{L} + k \log N \tag{14}$$

where $\log \mathcal{L}$ is the log-likelihood, $k$ is the number of model parameters, and $N$ is the number of observations, i.e., the number of gaze fixations for a given menu configuration. It is common practice to consider strong evidence in favor of a winning model when the difference between BIC scores is greater than 6 [55].

## 5. Results

To sum up, our results confirm that the Maximal model significantly outperforms the other models regardless the conditions (menu organization and menu length). However, none of the other models, in particular the Fluid and the Memory models, can describe well the participants data. Focusing on the best model, the Maximal model, the parameter values are consistent with what is known in the research literature (e.g., uncertainty about the landing point, distance performed by a saccade in serial search, or systematic bias in recall search), which is important to test the practical validity of the model. The results also confirm that the choice of visual search strategy evolves with practice. Typically, the proportion of Recall search linearly increases with practice while the proportion of Serial search decreases with practice. The proportion of Random search remains the same. Moreover menu organization influences the evolution of visual search strategies. For example, the decrease of serial search is more prominent for the Unordered organization. Finally, they show that the simulated gaze fixations reflect well participants' data for each visual search strategy both qualitatively and quantitatively. For example, we observe several ups and downs for Random search. The top-down inspection is well reflected with serial search. The number of fixations for recall search is smaller than for serial and random search (because of the direct access to the target item). In the following, we go over the evidence for these results. We compare and contrast the different arbitration mechanisms. We then offer detailed analysis of visual search strategies, including qualitative and quantitative analyses of gaze fixations.

Table 3: Comparisons of the arbitrators in term of free parameters, total number of free parameters (N), Likelihood, and BIC.

| **Model** | **Free parameters** | **N** | **log $\mathcal{L}$** | **BIC** |
|---|---|---|---|---|
| Serial-only | $A_{\text{serial}}$, $\sigma_{\text{serial}}$ | 2 | -435.2 | 879.3 |
| Recall-only | $B_{\text{recall}}$, $\sigma_{\text{recall}}$ | 2 | -406.8 | 822.3 |
| Random-only | $V_{\max}$, $\beta$ | 2 | -436.5 | 881.9 |
| Random | $A_{\text{serial}}$, $\sigma_{\text{serial}}$, $B_{\text{recall}}$, $\sigma_{\text{recall}}$, $V_{\max}$, $\beta$ | 6 | -404.4 | 835.2 |
| Fluid | $A_{\text{serial}}$, $\sigma_{\text{serial}}$, $B_{\text{recall}}$, $\sigma_{\text{recall}}$, $a$, $b$ | 6 | -406.9 | 840.3 |
| Memory | $A_{\text{serial}}$, $\sigma_{\text{serial}}$, $B_{\text{recall}}$, $\sigma_{\text{recall}}$, $a$, $b$, $\rho$ | 7 | -401.9 | 834.6 |
| Maximal | $A_{\text{serial}}$, $\sigma_{\text{serial}}$, $B_{\text{recall}}$, $\sigma_{\text{recall}}$, $V_{\max}$, $\beta$ | 6 | **-317.9** | **662.3** |

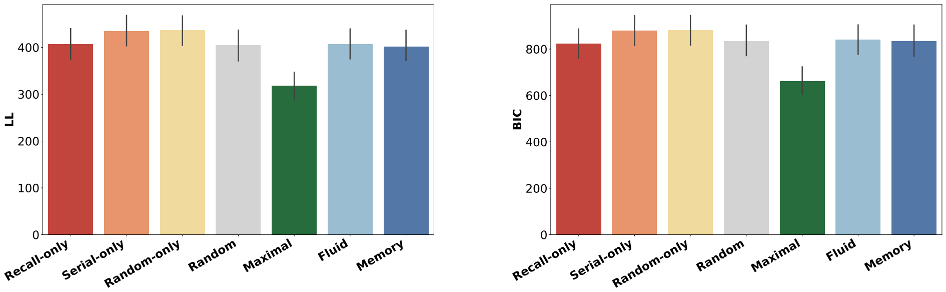


Figure 4: Model comparisons according to Log-likelihood (LL, left plot) and BIC score (right plot). The lower the better. Error bars show 95% confidence intervals.

### *5.1. Analysis of arbitration mechanisms*

Table 3 and Figure 4 indicate the likelihood and BIC scores for all arbitration mechanisms, regardless of the menu conditions. As expected, there is strong evidence (BIC $> 6$) that Maximal ($\log \mathcal{L} = -317.9$, BIC $= 662.3$) outperforms the other models, even when considering the penalty associated to the BIC score for additional model parameters. The second best model is Recall-only ($\log \mathcal{L} = -406.8$, BIC $= 822.3$). Finally, none of the other arbitration mechanisms (Serial-Only, Random-only, Fluid, Memory) significantly outperforms Random ($\log \mathcal{L} = -404.4$, BIC $= 835.2$).

We now refine our analysis by comparing the arbitration mechanisms for each menu condition (3 organizations × 3 lengths). The results are illus-

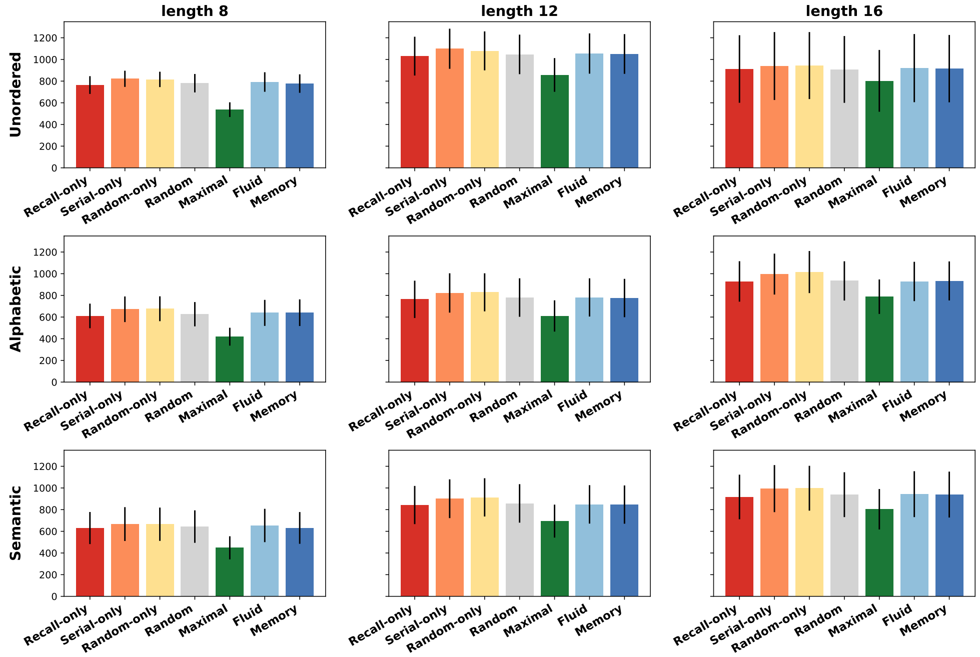


Figure 5: Model comparisons in terms of BIC score (lower is better) for the 3 menu organizations (Unordered, Alphabetic, Semantic) and 3 menu lengths (8, 12, 16). Error bars show 95% confidence intervals.

trated in Figure 5. We also ran an ANOVA test (Greenhouse–Geisser corrected) to study the effect of Arbitration mechanisms, Organisation, and Length on BIC score (details provided in Appendix B). ANOVA confirmed the same trends per condition, but the relative performance among arbitration mechanisms tends to decrease when menu length increases.

### *5.2. Analysis of visual search strategies*

We first analyze each visual search independently. We study the influence of menu length and menu organization on the values of each model's parameters. The analyses rely on the Maximal arbitration mechanism, designed to return the probability of each visual search strategy per trial and per participant. We consider only the subset of participants who have enough data for all conditions, i.e., 9 participants.

*Serial search.* The serial search strategy has two parameters: $A_{\text{serial}}$ and $\sigma_{\text{serial}}$. The average distance between two consecutive fixations ($A_{\text{serial}}$) is 3.4

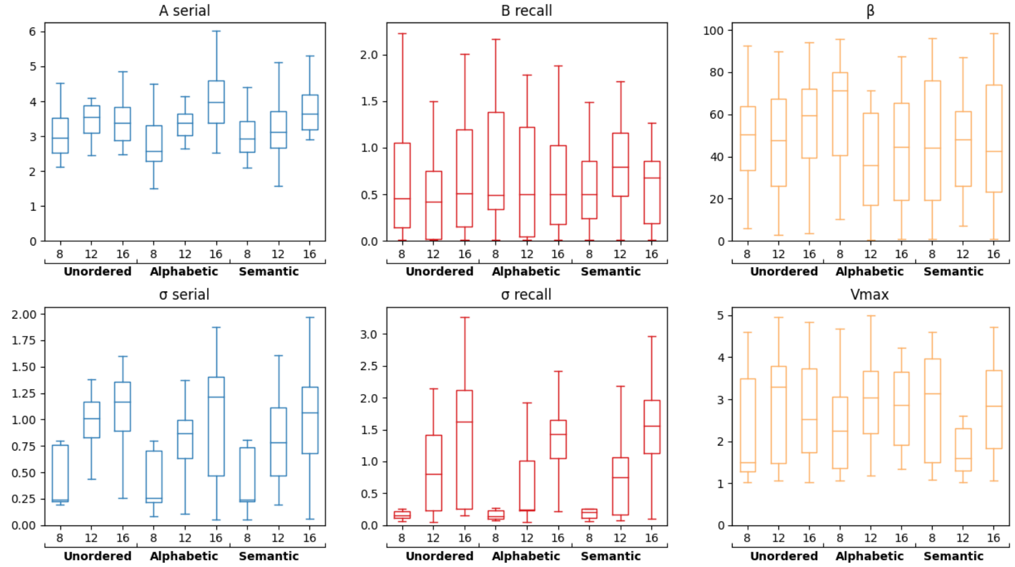


Figure 6: Analysis of model parameters (2 parameter per model) for each menu configuration and each visual search strategy. Leftmost column: serial search; middle column: recall search; rightmost column: random search.

items (std=1.0) which is in line with what was expected (see Section 3.2.1). Indeed, given the experimental setting, participants could focus on about three items at once if we consider a fovea of 2 degrees, as hinted in Section 4.1. ANOVA only reveals a significant effect of MENU LENGTH on $A_{\text{serial}}$ ($F_{2,16} = 12.7, p < .0001, \eta_p^2 = 0.18$) indicating that $A_{\text{serial}}$ increases with LENGTH (8 items: 3.1; 12 items: 3.5; 16 items: 3.8). The uncertainty of the landing point, $\sigma_{\text{serial}}$, is small in comparison with $A_{\text{serial}}$ with an average value of 0.8 (std=0.5). ANOVA indicates no significant effect of LENGTH or ORGANISATION on $\sigma_{\text{serial}}$.

*Recall search.* The Recall search has two parameters: $B_{\text{recall}}$ and $\sigma_{\text{recall}}$. The results indicate that the bias $B_{\text{recall}}$ is 0.7 items (std=0.6). This value is positive, confirming that users tend to systematically undershoot [27] the visual target in the recall search strategy. This value is also smaller than 3, confirming that the fixation landing remains in the fovea. ANOVA reveals no effect of ORGANIZATION or LENGTH on $B_{\text{recall}}$. The average value of $\sigma_{\text{recall}}$ is 0.8 (std=0.7). Similarly, to the serial search strategy, ANOVA indicates no significant effect of ORGANISATION on $\sigma_{\text{recall}}$, but it shows that the uncer-

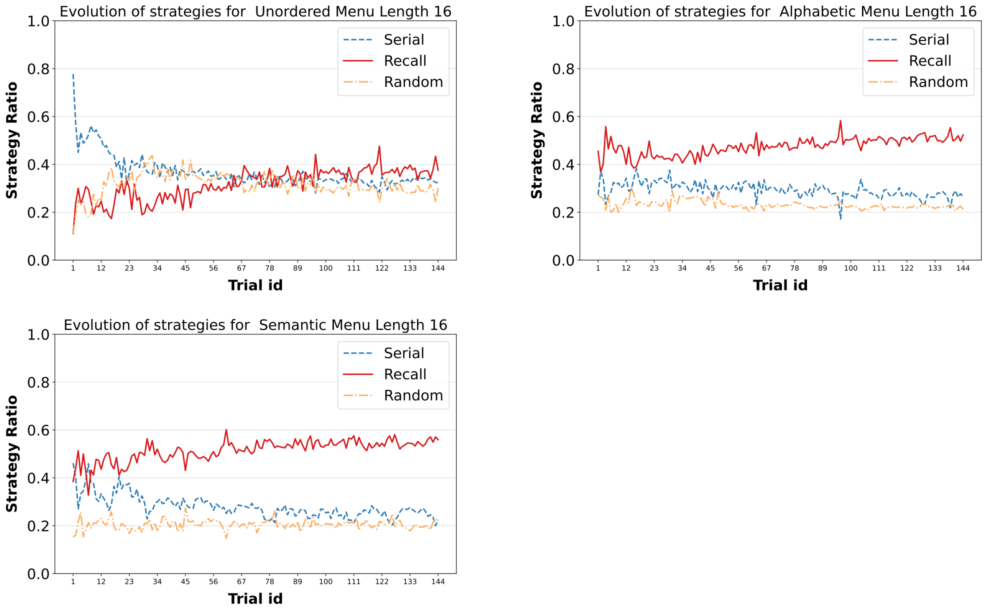


Figure 7: Proportion of the three visual search strategies as a function of trial id (earlier trials happened earlier in time) and menu organization for menus of length 16.

tainty significantly increases with menu length. Note that the value of $\sigma_{\text{serial}}$ and $\sigma_{\text{recall}}$ are similar (both of 0.8 items) while the average distance between two consecutive fixations is very different.

*Random search.* The Random search has two parameters: $V_{max}$ and $\beta$. The average number of items that users can remember, $V_{max}$ is 2.5 (std=1.2). ANOVA does not reveal a significant effect of Organization or Length on $V_{max}$. The average value of $\beta = 47.6$ is very high, indicating that an item which has been fixated is unlikely to be revisited again (random search without replacement). ANOVA does not reveal a significant effect of Organization or Length on $\beta$.

### *5.3. Proportion of visual search strategies*

To understand how visual search strategies evolve with practice, we rely again on the Maximal arbitration mechanism. For this analysis, we consider only menus of length 16, which is the most informative design and the more likely to highlight the differences between visual search strategies. For example, on an 8-items menu, if the target item is the second one, it is not possible to clearly distinguish a recall search from a serial search, considering the values of the model parameters discussed in the previous section. On menus with more items, this effect will diminish as the target item position increases.

Figure 7 shows the proportion of the three visual search strategies per trial and per menu organisation. We observe some similarities among menu organisations. Notably, the proportion of recall search increases with practice, with a percentage change of 20 between the beginning and the end of the experiment. The proportion of random search (25%) is almost systematically below the proportion of serial search (35%).

We also observe some differences among organizations: The initial proportion of recall search varies from 12% for Unordered to 45% for Alphabetic (40% for Semantic). The decrease of serial search is more prominent for Unordered (-45) than Semantic (-25) and Alphabetic (-5) menus. Finally, the proportion of random search is stable over trial for Alphabetic and Semantic menus, around 20%. In contrast, it increases until trial 30 and then decreases for Unordered menus.

Finally, we observe that the most frequent strategy rarely exceeds 60%, even when analyzing the data per participant (see Figure 6). This finding explains why the three X-only models do not explain well users' behavior; they can, at best, explain 60% of the observed data.

To sum up, these results confirm that the choice of visual search strategy evolves with practice and depends on the menu organisation.

### *5.4. Qualitative analysis of gaze fixations*

We refine our previous analysis by studying the sequence of fixated items generated by the Maximal arbitration mechanism, as it is the best model that explains the observed data. Figure 8 illustrates different sequences of fixated items (green) as a function of visual search strategies in a menu with 16 items. The Y axis indicates the vertical coordinate of each gaze fixation. The X axis indicates time, in seconds. The location of the target item is represented in a yellow color.

The first row shows both the sequence of gaze fixations (green) and the gaze path (red) for 2 users in the dataset (S18 and S19) chosen at random. The rows below provide simulation examples of search behaviors based on unseen trials. We observe that simulated data reflect well participants' data for each visual search strategy. Typically, we observe several ups and downs for random search. The top-down inspection is well reflected with serial search. The number of fixations for recall search is smaller with a direct access to the target item. We also observe in Figure 8 the stochasticity of our model: two simulations executed on the same configuration do not necessarily produce the same sequence of fixated items.

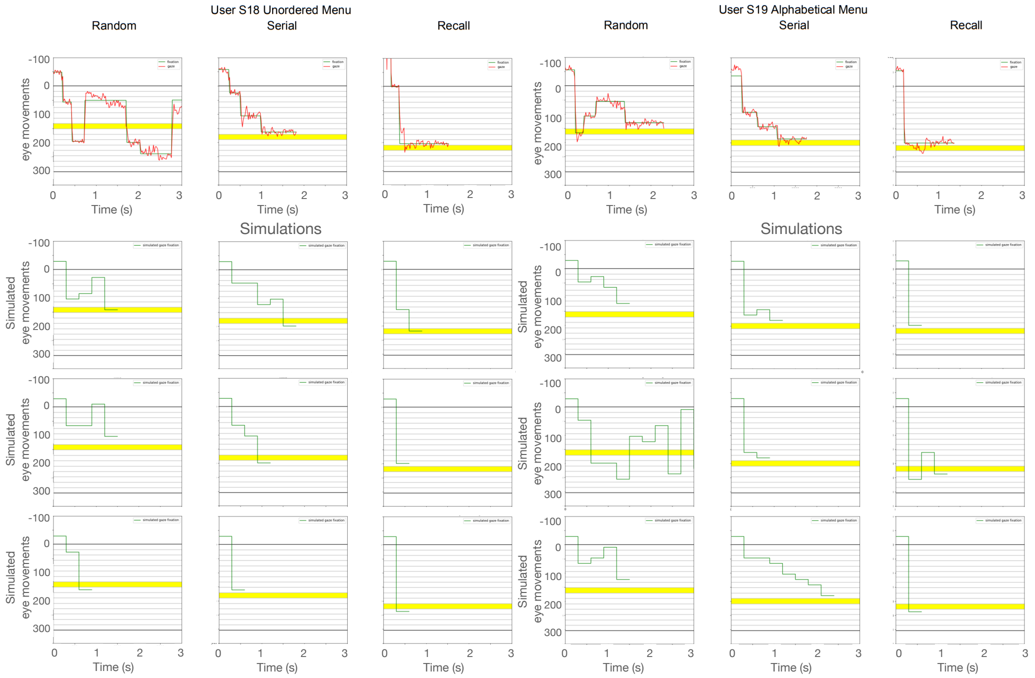


Figure 8: Models simulations of gaze fixations (green) for two users on menus with 16 items. The location of the target item is represented in yellow color. The top row shows actual user data, including saccadic eye movements (gaze path) in red. Left column: Unordered menus. Right column: Alphabetic menus. More examples of scanpaths from actual users are available at https://hci.isir.upmc.fr/project/menu-search-strategies/

### 5.5. *Quantitative analysis of gaze fixations*

A qualitative analysis of the sequences of fixated items is important to understand the output of the model, but it remains difficult to analyze in a quantitative way. For this reason, we analyze the (i) the total number of gaze fixations and (ii) the distribution of gaze fixations per item [8].

For a given participant, the model is trained on all blocks except the last 3 blocks. Then, we run the model in simulation mode (with the learned parameters) on each trial of the last 3 blocks (totaling 3,987 unseen trials in total) for 10 times. Finally, we quantitatively compare the **total number of gaze fixations** estimated from the simulated data ($\hat{\theta}_i$) and from human data ($\theta_i$). We report Mean Biased Error (MBE) and the Root Mean Squared Error (RMSE):[4]

$$\text{MBE} = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\theta}_i - \theta_i \right) \tag{15}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \hat{\theta}_i - \theta_i \right)^2} \tag{16}$$

Table 4 summarizes the results, averaged across all users in the dataset. As observed, the larger the number of items in the menu, the larger the MBE and RMSE values of the models. This finding is independent of the menu organization. The lowest values of both metrics were observed in Semantic menus with 8 items.

In addition, we can see that the model is never off by more than 2 (at most 3) fixations, as reported by the RMSE values in Table 4. This suggests that, while we observed some variation in the magnitude of the errors, it is unlikely that large errors have occurred. Further, according to the observed MBE values, given that the sign is always positive in Table 4, we can see that the model tends to overestimate the number of fixations. Therefore we can conclude that the Maximal model generalizes well to unseen trials, given previous user exposure to these menu configurations.

Finally, we refine this analysis and study the distribution of gaze fixations using the same method. Figure 9 shows the number of gaze fixations as a function of the location of the target item for both the observed human data and the simulated data on unseen trials. The nine plots reflect the 3 menu organizations (by column) × the 3 menu lengths (by row). This representation shares similarities with the one used by Chen et al. [8] and provides more information than Table 4. It shows that the simulated data

[4] Both MBE and RMSE measure the average magnitude of prediction errors. On the one hand, MBE considers the signed differences between predictions and actual observations. On the other hand, RMSE measures the root of the average of squared differences between predictions and actual observations. RMSE penalizes large errors more than MBE, therefore if large errors are undesirable, we should pay more attention to this metric.

| **Organization** | **Length** | **MBE** | **RMSE** |
|---|---|---|---|
| Unordered | 8 | 0.71 | 1.24 |
| | 12 | 0.88 | 1.98 |
| | 16 | 1.43 | 2.93 |
| Alphabetic | 8 | 0.74 | 1.16 |
| | 12 | 0.78 | 1.56 |
| | 16 | 1.23 | 2.93 |
| Semantic | 8 | 0.53 | 0.88 |
| | 12 | 0.94 | 1.83 |
| | 16 | 0.97 | 3.04 |

Table 4: Comparison of the total number of gaze fixations from simulated data and human data. The simulated data are generated on the last 3 blocks of the dataset, which represent unseen data for the model, by running 10 simulations per trial.

well reflect the observed human data regardless the location of the target item.

## 6. Discussion

### *6.1. The implementation of the visual search strategies is psychologically plausible*

There is often a tension between the capacity of a computational model to describe well the empirical data (descriptive adequacy) and the *plausibility* and *interpretability* of its parameters, i.e., whether the model parameters make sense and are psychologically valid [31]. For example, some approaches such as ABC [35] aim to find the best compromise between maximizing the fitness function and returning psychologically plausible values.

Our implementation of the three visual search strategies are grounded in HCI and psychology literature. However, it does not make assumptions about the parameter values which result from an optimization process without constraints. The parameter values are those which maximize the fitness function $\mathcal{L}_{\text{global}}$. Thus our approach, by definition, favors descriptive adequacy regardless of the model. One might then question whether the estimated parameter values are consistent with what is known in the research literature, which is important to test the practical validity of a model. In this regard, our results show that the parameter values of the Maximal model

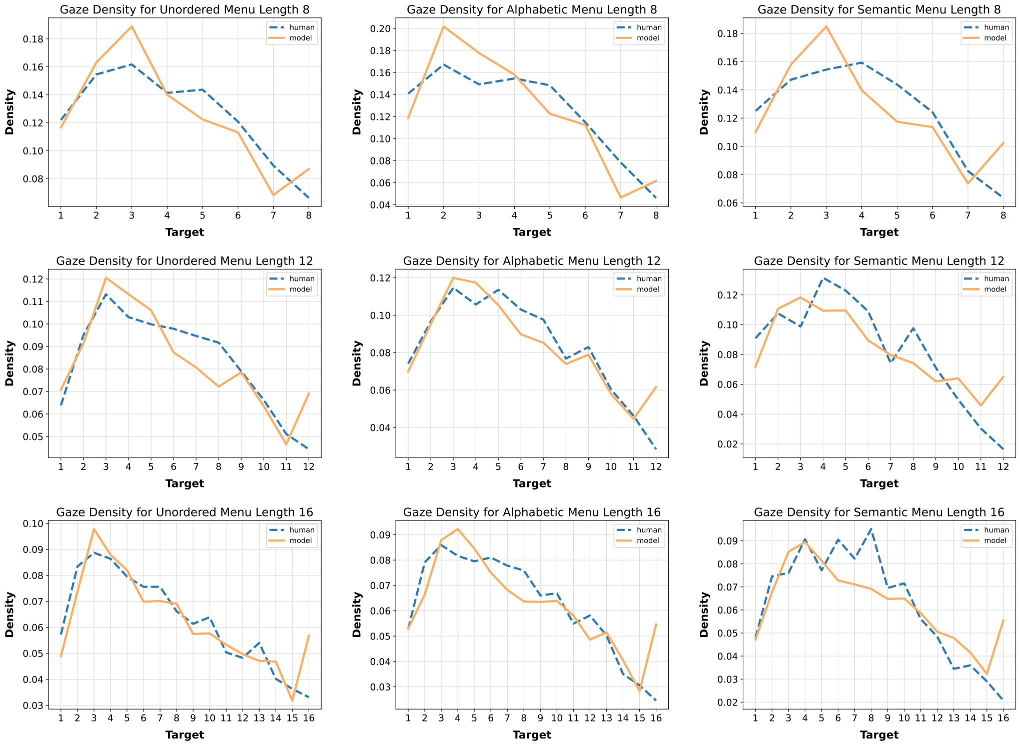


Figure 9: Distribution of gaze fixation per target item according to the observed users' search strategies (blue) and the simulated search strategies (orange). Each row is one of the menu lengths (8, 12, and 16 items). Each column is one of the menu organizations (Unordered, Alphabetic, and Semantic).

are psychologically plausible and consistent with the state-of-the art. In the following we summarize our key findings:

- The average distance between two fixations $A_{\text{serial}}$ is in the same range as the fovea distance, confirming that (i) users do not skip items when performing the serial search strategy and (ii) do not skim twice the same item when performing a top-down inspection, i.e., the field of vision associated with two consecutive gaze fixations do not overlap.

- The value of the bias $B_{\text{recall}}$ (0.7 items) confirms that users tend to systematically undershoot the target item in the recall search strategy, which is in line with [27]. Moreover, $B_{\text{recall}}$ is small enough that the target item is in the field of vision centered on the fixation landing point.

- The uncertainty of the fixation landing points $\sigma$ is smaller than the fovea distance and is independent of the visual search strategy. Indeed, it is likely that this uncertainty is a characteristic of the occulomotor system rather than a consequence of the task or strategy.

- The average number of menu items that users can remember is $V_{\max} = 2.5$.

- As discussed in the two previous sections, the whole dynamics of visual search strategies make sense given the characteristics of the menu organization and what we know about users' expertise.

In summary, despite the fact that the parameter values of the Maximal model are those that best describe the empirical data, they can also be interpreted (by definition of our model). More interestingly, they are also psychologically plausible. These results are promising and also interesting according to the computational rationality principle [22, 43]. Computational rationality poses that users make their behavioral choices governed by an attempt to maximize their utility which is bounded by their own resources (perceptive, motor and cognitive), their own experience, the constraints of the task (select the item as fast as possible), and the constraints of the environment (e.g., menu layout). In other words, the underlying idea of computational rationality is that the parameter values which are the most psychologically plausible are also those that best describe users' behavior. However, we have observed that it is not necessary to assume that users are optimal at every step in their behavior. Rather, it is enough to assume that they are good at picking a lower-level heuristic strategy to follow. This is also cognitively less involved, as such heuristics can be executed without resorting to planning.

### *6.2. The complexity of the models*

There is also a tension between model complexity and other criteria such as descriptive adequacy or plausibility. Currently, each of the three low-level mechanisms can be described with a few lines of code and have only two parameters. Similarly, the high-level mechanisms require a few more lines of code and more parameters. Moreover, the hierarchical structure of our approach splits the models in sub-models, thereby facilitating their comprehension, their reuse, and their extensibility. Most of the complexity lies in the method to evaluate the models and estimate their parameters, which is independent of the model implementations and is common to other cognitive evaluation methods [57].

### *6.3. Studying visual search strategies that are not observable*

Empirical studies on menu interaction primarily focus on selection time or gaze fixation which are directly observable with an eye-tracker. In contrast, visual search strategies are not directly observable: Given a sequence of gaze fixations, it requires interpretation to derive a visual search strategy. Several of them have been identified in the literature, but they have never been precisely/empirically quantified. As a result, existing computational models of visual search in linear menus make strong assumptions about the development of visual search strategies.

Our work contributes a sound methodology to the study of visual search strategies in menus and in GUIs by alleviating assumptions about visual search strategies. Our method requires as input a set of visual search strategies and the sequences of gaze fixations. Our Maximal model then returns the sequence of visual search strategies. We demonstrated that our method makes it possible to identify the strategy which best explains a user' trial data regardless of the user practice.

Reconstructing the visual search strategy at each trial offers several advantages. It augments our understanding of how users search in menus. It also enriches existing data set about visual search in menus to allow HCI researchers to test different hypotheses. For example, we augmented the dataset of Bailly et al. [5] with the probability of each visual search strategy to be adopted at each trial. The new dataset will be publicly available and can be used to elaborate and test novel computational models of visual search. The dataset is currently here: link

### *6.4. Learning impacts the choice of visual search strategies*

Our results first reveal that the X-only arbitrators do not describe well the empirical data, i.e., a single strategy does not explain well the diversity in users' behavior. In contrast, the Maximal arbitrator better explains users' behavior and reveals that each individual adopts several visual search strategies over time. Further analysis shows that the proportion of these strategies evolve with users' experience. In particular the proportion of recall search linearly increases with practice. However, the tendency is not as prominent as expected regardless of menu organisation. Indeed, even after 9 blocks of training, the proportion of recall search remains lower than 60%. These results suggest than users did not fully transition to expert behavior. These results are in line with previous findings about keyboard and gesture shortcuts adoption where, despite a lot of training, users did not (fully) transition to expert behavior [25, 40, 61]. The proportion of serial search decreases with practice while the proportion of random search remains the same. This result confirms that the transition from novice to expert behavior is primarily explained by the transition from serial to recall search.

These results also explain the low performance of the Fluid and Memory arbitrators. These models have the advantage of describing the transition from novice to expert behavior in a simple way. In particular, they did not account for random search because we were expecting a low proportion of this strategy. Our results show quite the opposite. The random search strategy represents 20% of the selection. Moreover its dynamics are subtle, with an inverted U shaped curve. Finally, the performance of the random model is higher than expected, i.e., higher than the Fluid and the Memory models. This is explained by the fact that the proportion of the three strategies always remain between 20% and 60%, which is not so different from the 33% probability of the random model.

### *6.5. Menu organization and length influence visual search strategies*

*High level cognitive processing.* Our results show an impact of menu organisation on the initial proportion of recall search. Users have the highest initial proportion of recall search (45%) with the alphabetic menu organisation. It can be explained by the nature of the stimulus of the experiment. Indeed, if the stimulus is the name of the command to execute, users can then rely on their knowledge of the alphabet to guess where to look at. For example, the item "Save" is likely to be in the bottom part of the menu in a alphabetic organisation as the first letter is 'S'. In contrast, the Unordered menu organisation has the lowest initial proportion of recall search (12%) because users have no clear cues to guess where the target item is located from the command name. Finally, in Semantic menus, novice users can quickly learn to ignore some groups[5] and jump directly to the right one.

[5] They ignore groups by considering both the semantic proximity between the items of a same group and the semantic distance between the stimulus and the fixated item.

*Low level cognitive processing.* We have observed that the organisation of the menu or its length have no significant impact on the values of the different parameters of the visual search strategies. One exception is the distance between two consecutive fixations $A_{\text{serial}}$ in the serial search strategy. This distance slightly augments with menu length, $+0.09\,\text{cm}$ per every additional item. It can be explained by the fact that users take more risks (more likely to skip an item) when the length of the menu increases.

### *6.6. Limitations and Future work*

Our research opens up new directions for future work at different fronts, which we discuss as follows.

*Visual search strategies.* One assumption of our models is that user's practice does not influence each visual search strategy. This assumption seems reasonable for recall search. This is less clear for serial and random search. For example, it would be interesting to study whether user's practice has an impact on the distance between two consecutive fixations, $A_{\text{Serial}}$ in serial search. More generally, future work should investigate the impact of user's practice on each parameter. Moreover, our implementation of serial search is limited to top-down inspection and could be extended to bottom-up inspection. Certain variants of the random search strategy could also be considered, such as one introducing a bias toward the top of the menu. The uncertainty of the landing point in the serial and recall strategy is represented by a Gaussian distribution. Future work should consider more advanced distribution. Another assumption of our models is that users adopt a unique strategy when searching for a target item. However, a user can start with a recall search, not find the desired item and move forward with a serial search. Future work should investigate the benefits of combining different strategies when searching for a given target item. Moreover, it might be useful to consider additional search strategies like semantic-based visual search strategies in hierarchical structures [60] such as foraging search [62]. More generally, future work should consider the variety of real-world menus such as radial menus or grid-based menus, which may engage different cognitive processes Finally, future work should study what is really specific to a given visual search strategy and what is common to the human oculomotor system. For example, both serial and recall have a parameter describing the uncertainty about the landing point $(\sigma_{serial}, \sigma_{recall})$ which are currently independent. Our results suggest that a unique parameter reflecting the uncertainty of the fixation landing may be sufficient.

*Arbitration mechanisms.* While our results are promising regarding the description of each visual search strategy, they are less promising regarding the description of arbitration mechanisms. This may be explained by the fact that previous work focused much more on the visual search strategies (low-level processes) than the mechanisms to choose them (high-level processes).

Future work should more deeply understand why and how users choose search strategies. We underestimated the proportion of random search and we did not anticipate a linear increase of recall search over time. Future work should confirm these phenomena as well as new models to cover them. A promising direction for such models would be to introduce an Explore/Exploit component to refine the arbitration mechanisms. Similar to Bailly et al. [2], the component can update the value of each visual search strategy according to the time needed to find the target item. The component would have the advantage to not be limited to two strategies.

*Data collection and long-term learning effects.* As previously discussed, our results show that participants did not systematically adopt recall search by the end of the experiment. This suggests that they did not really become experts for the given number of trials. Moreover, the data collection only contains menu with 8–16 items. Future work should consider collecting data (1) with longer menus (longer are the menus, easier it is to distinguish visual search strategies) and (2) over longer periods, in order to let more time to participants to get more familiar with the menus.

*Implications for design.* Our work contributes a better understanding of users' search behavior with GUIs and menus in particular. It also contributes effort towards the elaboration of efficient model-based adaptive menus. A main challenge of this class of interfaces is to diminish the short-term cost of picking an adaptation. Designers can build on our empirical findings and models to help users find the novel location of a target item. For example, consider that the best possible location of an item is at the top of the menu (as it is faster to reach) but it is currently placed at the bottom. If the user is likely to perform a serial search, the system can immediately move the item to the top. However, if the user performs a recall search, it might be more beneficial to move up the item progressively, i.e., move it up one item every

$k$ trials so that the landing point of recall search (which tends to undershoot) remains in the vicinity of the target item.

*Beyond visual search.* Finally, another promising direction for future work is to extend our methodology to different use cases. A compelling use case, for example, is to investigate how users choose pointing strategies in menus [5, 7, 14] and how they evolve with practice. One strategy consists of *moving* the cursor once the target item has been located by visual search. Another one is *tracking* where the cursor follows the gaze. Finally, in a *tagging* strategy, the cursor is used to tag an item while the eyes are free to move. An interesting challenge is to explain the interaction effects between pointing strategies and visual search strategies.

## 7. Conclusion

We have studied different models of visual search strategies in menus that can explain how visual search evolves with experience. While the search strategies are not directly observable, our hierarchical arbitration mechanism allows to automatically infer them from a sequence of gaze fixations. Our findings also suggest that the proportion of serial and recall search evolves with practice and depends on the menu organization, however this novice-to-expert transition is not as apparent as the research literature would have expected, even after multiple exposures. Finding an item in a menu is a common task in HCI that involves subtle learning and decision-making mechanisms, therefore our findings can inform researchers and other domain areas well beyond menus. Our software and derivative data will be made publicly available upon publication.

## Acknowledgments

We thank Benoit Girard for his useful feedback. This work was supported by the Department of Communications and Networking – Aalto University, the Finnish Center for Artificial Intelligence (FCAI), the ANR NeuroHCI project (ANR-22-CE33-0006-01), the Academy of Finland through the Human Automata (grant 328813) and BAD (grant 318559) projects, the Horizon 2020 FET program of the European Union (grant CHIST-ERA-20-BCI-001) and the European Innovation Council Pathfinder program (grant 101071147).

## References

[1] Anderson, J.R., Lebiere, C.J., 2014. The atomic components of thought. Psychology Press. doi:10.4324/9781315805696.

[2] Bailly, G., Khamassi, M., Girard, B., 2022. Computational model of the transition from novice to expert interaction techniques. ACM Trans. Comput.-Hum. Interact. URL: https://doi.org/10.1145/3505557, doi:10.1145/3505557. just Accepted.

[3] Bailly, G., Lecolinet, E., Nigay, L., 2016. Visual menu techniques. ACM Comput. Surv. 49. URL: https://doi.org/10.1145/3002171, doi:10.1145/3002171.

[4] Bailly, G., Malacria, S., 2022. Command selection, in: Vanderdonckt, J., Palanque, P., Winckler, M. (Eds.), Handbook of Human Computer Interaction. Springer International Publishing, Cham, pp. 1–35. URL: https://doi.org/10.1007/978-3-319-27648-9_19-1, doi:10.1007/978-3-319-27648-9_19-1.

[5] Bailly, G., Oulasvirta, A., Brumby, D.P., Howes, A., 2014. Model of visual search and selection time in linear menus, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA. pp. 3865–3874. URL: http://doi.acm.org/10.1145/2556288.2557093, doi:10.1145/2556288.2557093.

[6] Bailly, G., Oulasvirta, A., Kötzing, T., Hoppe, S., 2013. Menuoptimizer: Interactive optimization of menu systems, in: Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, Association for Computing Machinery, New York, NY, USA. pp. 331–342. URL: https://doi.org/10.1145/2501988.2502024, doi:10.1145/2501988.2502024.

[7] Byrne, M.D., Anderson, J.R., Douglass, S., Matessa, M., 1999. Eye tracking the visual search of click-down menus, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 402–409. URL: https://doi.org/10.1145/302979.303118, doi:10.1145/302979.303118.

[8] Chen, X., Bailly, G., Brumby, D.P., Oulasvirta, A., Howes, A., 2015. The emergence of interactive behavior: A model of rational menu search, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY,

USA. p. 4217–4226. URL: https://doi.org/10.1145/2702123.2702483, doi:10.1145/2702123.2702483.

[9] Cinotti, F., Fresno, V., Aklil, N., Coutureau, E., Girard, B., Marchand, A.R., Khamassi, M., 2019. Dopamine blockade impairs the exploration-exploitation trade-off in rats. Scientific reports 9, 1–14.

[10] Clarke, A.D., Stainer, M.J., Tatler, B.W., Hunt, A.R., 2017. The saccadic flow baseline: Accounting for image-independent biases in fixation behavior. Journal of vision 17, 12–12.

[11] Cockburn, A., Gutwin, C., 2009. A predictive model of human performance with scrolling and hierarchical lists. Human–Computer Interaction 24, 273–314. doi:10.1080/07370020902990402, arXiv:https://www.tandfonline.com/doi/pdf/10.1080/07370020902990402.

[12] Cockburn, A., Gutwin, C., Greenberg, S., 2007. A predictive model of menu performance, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA. pp. 627–636. URL: http://doi.acm.org/10.1145/1240624.1240723, doi:10.1145/1240624.1240723.

[13] Coutrot, A., Hsiao, J.H., Chan, A.B., 2018. Scanpath modeling and classification with hidden markov models. Behavior research methods 50, 362–379.

[14] Cox, A.L., Silva, M.M., 2006. The role of mouse movements in interactive search, in: Proceedings of the Annual Meeting of the Cognitive Science Society.

[15] Daw, N., 2011. Trial-by-trial data analysis using computational models. Oxford University Press. doi:10.1093/acprof:oso/9780199600434.003.0001. publisher Copyright: © The International Association for the study of Attention and Performance, 2011. All rights reserved.

[16] Dayama, N.R., Shiripour, M., Oulasvirta, A., Ivanko, E., Karrenbauer, A., 2021. Foraging-based optimization of menu systems. International Journal of Human-Computer Studies 151, 102624. URL: https://www.sciencedirect.com/science/article/pii/S1071581921000422, doi:https://doi.org/10.1016/j.ijhcs.2021.102624.

[17] Delamare, W., Neshati, A., Irani, P., Ren, X., 2019. An analytic model for time efficient personal hierarchies, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing

Machinery, New York, NY, USA. p. 1–11. URL: https://doi.org/10.1145/3290605.3300598, doi:10.1145/3290605.3300598.

[18] Fahimi, R., Bruce, N.D., 2021. On metrics for measuring scanpath similarity. Behavior Research Methods 53, 609–628.

[19] Findlater, L., Moffatt, K., McGrenere, J., Dawson, J., 2009. Ephemeral adaptation: The use of gradual onset to improve menu selection performance, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. pp. 1655–1664. URL: https://doi.org/10.1145/1518701.1518956, doi:10.1145/1518701.1518956.

[20] Findlay, J.M., Gilchrist, I.D., 2003. Active vision: The psychology of looking and seeing. 37, Oxford University Press.

[21] Fitts, P.M., Posner, M.I., 1967. Human performance. Brooks/Cole.

[22] Gershman, S.J., 2020. Origin of perseveration in the trade-off between reward and complexity. Cognition 204, 104394.

[23] Goubko, M.V., Danilenko, A.I., 2010. An automated routine for menu structure optimization, in: Proceedings of the 2nd ACM SIGCHI symposium on Engineering interactive computing systems, pp. 67–76.

[24] Grossman, T., Dragicevic, P., Balakrishnan, R., 2007. Strategies for accelerating on-line learning of hotkeys, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA. pp. 1591–1600. URL: http://doi.acm.org/10.1145/1240624.1240865, doi:10.1145/1240624.1240865.

[25] Gutwin, C., Cockburn, A., Scarr, J., Malacria, S., Olson, S.C., 2014. Faster command selection on tablets with fasttap, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 2617–2626. URL: https://doi.org/10.1145/2556288.2557136, doi:10.1145/2556288.2557136.

[26] Halverson, T., Hornof, A.J., 2011. A computational model of “active vision” for visual search in human–computer interaction. Human–Computer Interaction 26, 285–314. URL: https://doi.org/10.1080/07370024.2011.625237, doi:10.1080/07370024.2011.625237, arXiv:https://doi.org/10.1080/07370024.2011.625237.

[27] Henson, D., 1978. Corrective saccades: Effects of altering visual feedback. Vision Research 18, 63–67. URL: https://www.sciencedirect.com/science/article/pii/0042698978900780, doi:https://doi.org/10.1016/0042-6989(78)90078-0.

[28] Hornof, A.J., 2004. Cognitive strategies for the visual search of hierarchical computer displays. Human–Computer Interaction 19, 183–223. doi:10.1207/s15327051hci1903_1.

[29] Hornof, A.J., Kieras, D.E., 1997. Cognitive modeling reveals menu search in both random and systematic, in: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 107–114. URL: https://doi.org/10.1145/258549.258621, doi:10.1145/258549.258621.

[30] Hu, Z., Bulling, A., Li, S., Wang, G., 2021. Fixationnet: Forecasting eye fixations in task-oriented virtual environments. IEEE Transactions on Visualization and Computer Graphics 27, 2681–2690.

[31] Jacobs, A.M., Grainger, J., 1994. Models of visual word recognition: sampling the state of the art. Journal of Experimental Psychology: Human perception and performance 20, 1311.

[32] Jiang, Y., Guo, Z., Rezazadegan Tavakoli, H., Leiva, L.A., Oulasvirta, A., 2024. Eyeformer: predicting personalized scanpaths with transformer-guided reinforcement learning, in: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, pp. 1–15.

[33] Jokinen, J.P., Wang, Z., Sarcar, S., Oulasvirta, A., Ren, X., 2020. Adaptive feature guidance: Modelling visual search with graphical layouts. International Journal of Human-Computer Studies 136, 102376. URL: https://www.sciencedirect.com/science/article/pii/S1071581919301429, doi:https://doi.org/10.1016/j.ijhcs.2019.102376.

[34] Kangasrääsiö, A., Athukorala, K., Howes, A., Corander, J., Kaski, S., Oulasvirta, A., 2017a. Inferring cognitive models from data using approximate bayesian computation, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 1295–1306. URL: https://doi.org/10.1145/3025453.3025576, doi:10.1145/3025453.3025576.

[35] Kangasrääsiö, A., Athukorala, K., Howes, A., Corander, J., Kaski, S., Oulasvirta, A., 2017b. Inferring cognitive models from data using approximate bayesian computation, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 1295–1306. URL: https://doi.org/10.1145/3025453.3025576, doi:10.1145/3025453.3025576.

[36] Klein, R.M., 2000. Inhibition of return. Trends in Cognitive Sciences 4, 138–147. URL: https://www.sciencedirect.com/science/article/pii/S1364661300014522, doi:https://doi.org/10.1016/S1364-6613(00)01452-2.

[37] Kümmerer, M., Bethge, M., 2021. State-of-the-art in human scanpath prediction. arXiv preprint arXiv:2102.12239 .

[38] Kümmerer, M., Wallis, T.S., Bethge, M., 2015. Information-theoretic model comparison unifies saliency metrics. Proceedings of the National Academy of Sciences 112, 16054–16059.

[39] Kunar, M.A., Flusberg, S., Wolfe, J.M., 2008. The role of memory and restricted context in repeated visual search. Perception & Psychophysics 70, 314–328.

[40] Lane, D.M., Napier, H.A., Peres, S.C., Sandor, A., 2005. Hidden Costs of Graphical User Interfaces: Failure to Make the Transition from Menus and Icon Toolbars to Keyboard Shortcuts. International Journal of Human-Computer Interaction 18, 133–144. URL: http://www.tandfonline.com/doi/abs/10.1207/s15327590ijhc1802_1, doi:10.1207/s15327590ijhc1802_1.

[41] Lee, E., MacGregor, J., 1985. Minimizing user search time in menu retrieval systems. Human Factors 27, 157–162.

[42] Lee, N., 2005. Interview with bill kinder: January 13, 2005. Comput. Entertain. 3. URL: http://doi.acm.org/10.1145/1057270.1057278, doi:10.1145/1057270.1057278.

[43] Lewis, R.L., Howes, A., Singh, S., 2014. Computational rationality: Linking mechanism and behavior through bounded utility maximization. Topics in cognitive science 6, 279–311.

[44] Li, Y., Bengio, S., Bailly, G., 2018. Predicting human performance in vertical menu selection using deep learning, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 1–7. URL: https://doi.org/10.1145/3173574.3173603, doi:10.1145/3173574.3173603.

[45] Li, Z., Ko, Y.J., Putkonen, A., Feiz, S., Ashok, V., Ramakrishnan, I., Oulasvirta, A., Bi, X., 2023a. Modeling touch-based menu selection performance of blind users via reinforcement learning, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. URL: https://doi.org/10.1145/3544548.3580640, doi:10.1145/3544548.3580640.

[46] Li, Z., Ko, Y.J., Putkonen, A., Feiz, S., Ashok, V., Ramakrishnan, I., Oulasvirta, A., Bi, X., 2023b. Modeling touch-based menu selection performance of blind users via reinforcement learning, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. URL: https://doi.org/10.1145/3544548.3580640, doi:10.1145/3544548.3580640.

[47] Matsui, S., Yamada, S., 2008. Genetic algorithm can optimize hierarchical menus, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 1385–1388. URL: https://doi.org/10.1145/1357054.1357271, doi:10.1145/1357054.1357271.

[48] McDonald, J.E., Stone, J.D., Liebelt, L.S., 1983. Searching for items in menus: The effects of organization and type of target, in: Proceedings of the human factors society annual meeting, SAGE Publications Sage CA: Los Angeles, CA. pp. 834–837.

[49] Molinaro, G., Collins, A.G., 2023. A goal-centric outlook on learning. Trends in Cognitive Sciences .

[50] Neisser, U., Beller, H.K., 1965. Searching through word lists. British Journal of Psychology 56, 349–358.

[51] Newell, A., Rosenbloom, P.S., 1993. Mechanisms of Skill Acquisition and the Law of Practice. MIT Press, Cambridge, MA, USA. p. 81–135.

[52] Nilsen, E.L., 1991. Perceptual-motor control in human-computer interaction. University of Michigan.

[53] Norman, K.L., 1991. The psychology of menu selection: Designing cognitive control at the human/computer interface. Intellect Books.

[54] Price, K.V., Storn, R.M., Lampinen, J.A., 2005. Differential Evolution: A Practical Approach to Global Optimization. 1st ed., Springer Berlin, Heidelberg.

[55] Raftery, A.E., 1995. Bayesian model selection in social research. Sociological methodology , 111–163.

[56] Rayner, K., 1998. Eye movements in reading and information processing: 20 years of research. Psychological bulletin 124, 372.

[57] Schütt, H.H., Rothkegel, L.O., Trukenbrod, H.A., Reich, S., Wichmann, F.A., Engbert, R., 2017. Likelihood-based parameter estimation and comparison of dynamical cognitive models. Psychological review 124, 505.

[58] Sears, A., Shneiderman, B., 1994. Split menus: Effectively using selection frequency to organize menus. ACM Trans. Comput.-Hum. Interact. 1, 27–51. URL: https://doi.org/10.1145/174630.174632, doi:10.1145/174630.174632.

[59] Somberg, B.L., 1986. A comparison of rule-based and positionally constant arrangements of computer menu items, in: Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface, Association for Computing Machinery, New York, NY, USA. p. 255–260. URL: https://doi.org/10.1145/29933.275639, doi:10.1145/29933.275639.

[60] Sourulahti, S., Janssen, C.P., Jokinen, J.P., 2024. Modeling rational adaptation of visual search to hierarchical structures. URL: https://arxiv.org/abs/2409.08967, arXiv:2409.08967.

[61] Tak, S., Westendorp, P., van Rooij, I., 2013. Satisficing and the use of keyboard shortcuts: Being good enough is enough? Interacting with Computers 25, 404–416. doi:10.1093/iwc/iwt016.

[62] Todi, K., Bailly, G., Leiva, L.A., Oulasvirta, A., 2021. Adapting user interfaces with model-based reinforcement learning. arXiv preprint arXiv:2103.06807 .

[63] Tsandilas, T., schraefel, m.c., 2007. Bubbling menus: A selective mechanism for accessing hierarchical drop-down menus, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association

for Computing Machinery, New York, NY, USA. p. 1195–1204. URL: https://doi.org/10.1145/1240624.1240806, doi:10.1145/1240624.1240806.

[64] Viejo, G., Khamassi, M., Brovelli, A., Girard, B., 2015. Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. Frontiers in behavioral neuroscience 9, 225.

[65] Wang, Y., Bâce, M., Bulling, A., 2023. Scanpath prediction on information visualisations. IEEE Transactions on Visualization and Computer Graphics 30, 3902–3914. URL: https://doi.org/10.1109/TVCG.2023.3242293, doi:10.1109/TVCG.2023.3242293.

[66] Wilson, R.C., Collins, A.G., 2019. Ten simple rules for the computational modeling of behavioral data. Elife 8, e49547.

[67] Wolfe, J.M., Klempen, N., Dahlen, K., 2000. Postattentive vision. Journal of Experimental Psychology: Human Perception and Performance 26, 693.

# Appendices

## Appendix A. Detailed formulation of the fitness function

The fitness function (or log-likelihood) is defined as:

$$\begin{aligned}\log \mathcal{L}(m, p; \theta) &= \log \prod_f P(a^p_f | a^p_1 ... a^p_{f-1}\}, m, p; \theta) \\ &= \sum_f \log P(a^p_f | a^p_1 ... a^p_{f-1}\}, m, p; \theta) \\ &= \sum_f \log P(a^p_f | S^p, m, p; \theta)\end{aligned} \tag{A.1}$$

where, $m$, $p$, $\theta$ are the model, the participant, and the set of parameters of the model $m$ for the participant $p$. $S^p$ is the whole sequence of fixated items from the beginning of the experiment. $a^p_f$ (low-level action) is the fth fixated item in this sequence. We refine this formulation by considering the sequence of fixated items per trial:

$$\log \mathcal{L}(m, p, \theta) = \sum_t \sum_f \log P(a^p_f | S^p, m, p; \theta) \tag{A.2}$$

where $t$ is the current trial index and $f$ is now the current gaze fixation index within each trial.

To make the choice of strategy $a'_t$ (high-level action) at the trial $t$ explicit, the loglikelihood function can be rewritten as:

$$\log \mathcal{L}(m, p, \theta) = \sum_t \sum_f \log \left( \sum_{a'} P(a'_t) P(a^p_f | S^p, a', m; \theta) \right) \tag{A.3}$$

where $P(a'_t)$ is the probability to choose one strategy $a'$ given the strategies chosen by the participant until the trial $t$. A key point here is the estimation of $P(a'_t)$ which is dependent of the arbitration mechanism. For the three X-only arbitrators (X $\in$ {serial,recall,random}) $P(a'_t) = 1$ if $a'_t = a'^X$ and 0 otherwise. For the Random arbitrator, $P(a'_t) = 0.33$ regardless the strategy. Finally, for the Maximal arbitrator, the idea is that $P(a'_t) = 1$ if $a'_t$ is the best strategy $a'_{\text{opt}}$ that best describes the sequence of gaze fixations at trial $t$ based on the data, and 0 otherwise. Formally,

$$a'_{\text{opt}} = \arg\min_{a'} \mathcal{L}_{\text{local}}(a') \tag{A.4}$$

where $\mathcal{L}_{\text{local}}(a')$ is the local fitness function:

$$\mathcal{L}_{\text{local}}(a') = \mathcal{L}_{\text{local}}(a', m, p, \theta) = \sum_{f} \log P(a_f^p | S^p, a', m; \theta) \tag{A.5}$$

## Appendix B. Model comparisons per condition

We ran an ANOVA test (Greenhouse–Geisser corrected) to study the effect of Model × Organisation × Length on BIC score. ANOVA only reveals a significant effect of Length ($F_{2,36} = 19122, p < .001, \eta_p^2 = 0.23$). A post-hoc tukey test shows that the models better describe menus with less items (8 items: 750; 12 items: 1,026; 16 items: 1,144) which is expected given the fact that there are less possible choices on small menus. Finally, ANOVA also reveals a significant Model × Length interaction effect ($F_{12,216} = 16, p < .001, \eta_p^2 = 0.002$) showing that the difference of relative performance among models tends to decrease when the length increases.

## Appendix C. Model comparisons per target item

Figure C.10 shows the BIC scores as a function of target item location and menu organisation for menus of length 16. The patterns are similar for shorter menus (8 and 12 items).

## Appendix D. Visual search strategies per user and organisation

Figure D.11 shows the proportion of the three visual search strategies as a function of user id and menu organization, for all menus of 16 items. The patterns are similar for shorter menus (8 and 12 items).

## Appendix E. Menu contents

*Length 8.*

- Set A: Desk, Chair, Coal, Couch, Gas, Oil, Table, Wood
- Set B: Artist, Banker, Boxing, Karate, Lawyer, Rugby, Teacher, Tennis
- Set C: Ears, Eye, Mail, Mouth, Nose, Phone, Radio, Skype

*Length 12.*

- Set A: Blender, Brain, Church, Coffeemaker, Heart, Juicer, Liver, Lung, Mosque, Shrine, Temple, Toaster.
- Set B: Bracelet, Canary, Carrot, Eagle, Earring, Lettuce, Necklace, Pigeon, Potato, Robin, Tomato, Watch
- Set C : Eraser, Eyeliner, Flute, Guitar, Lipstick, Marker, Mascara, Paper, Perfume, Pencil, Piano, Violin.

*Length 16.*

- Set A: Brake, Cancer, Cinema, Diabetes, Engine, Fork, Gears, Herpes, Knife, Leukemia, Museum, Plate, School, Spoon, Theater, Wheel
- Set B: Aunt, Brother, Canyon, Stove, Cousin, Diamond, Freezer, Emerald, Hill, River, Fridge, Ruby, Pearl, Dryer, Uncle, Valley
- Set C: Ballet, Butter, Cheese, Cream, Disco, Donkey, Horse, Jacket, Jazz, Pants, Rabbit, Sheep, Shirt, Socks, Tango, Yogurt

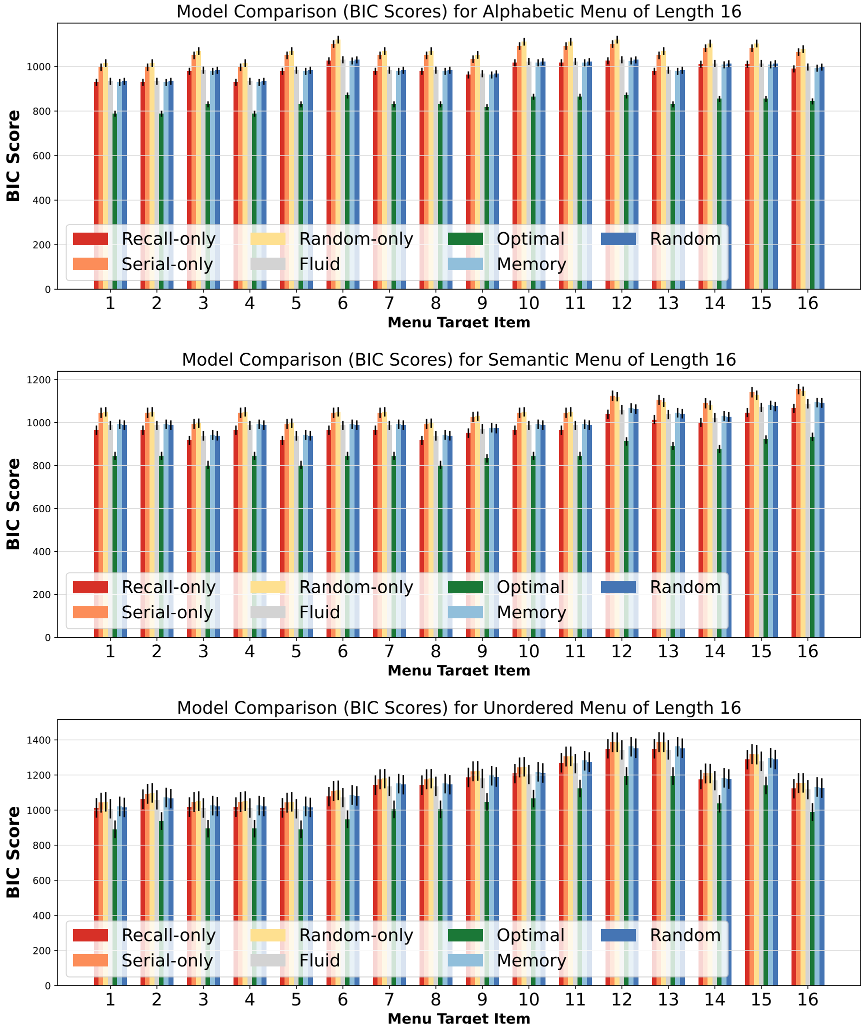


Figure C.10: Model comparisons (BIC score, lower is better) for the 16 positions of target items. Error bars show 95% confidence intervals.

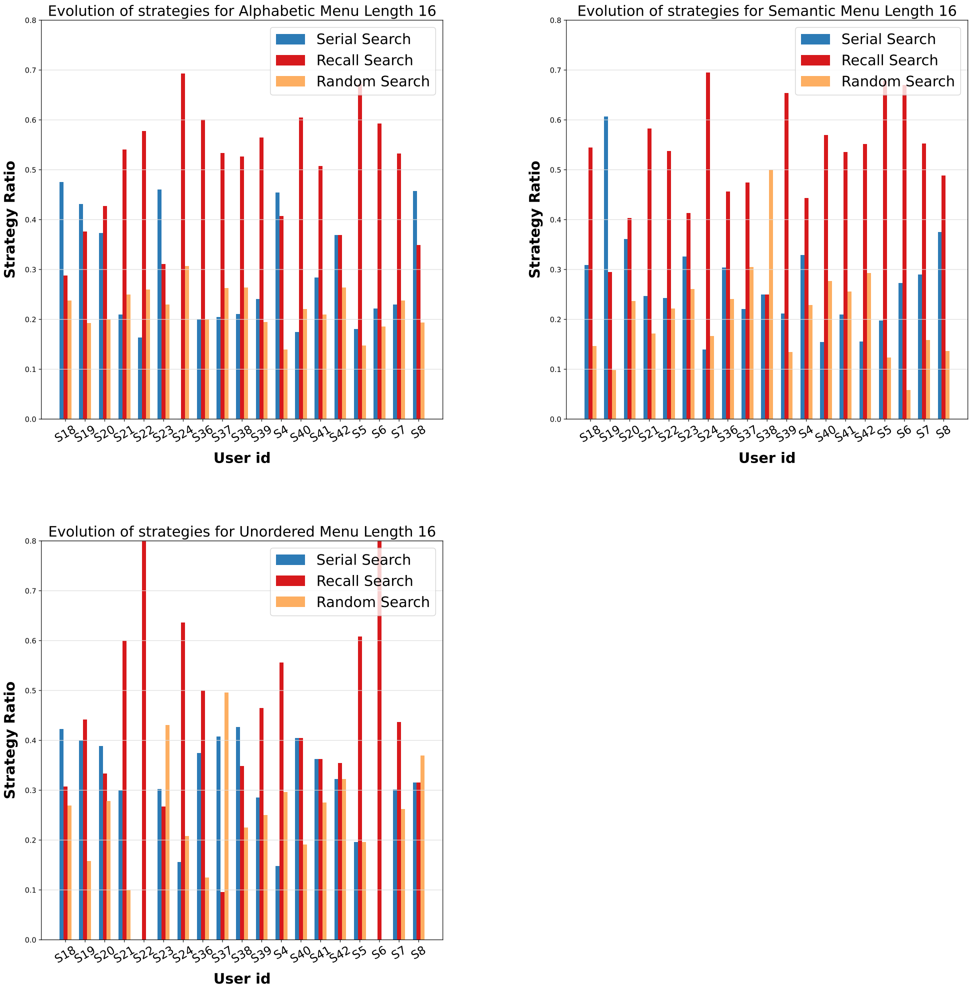
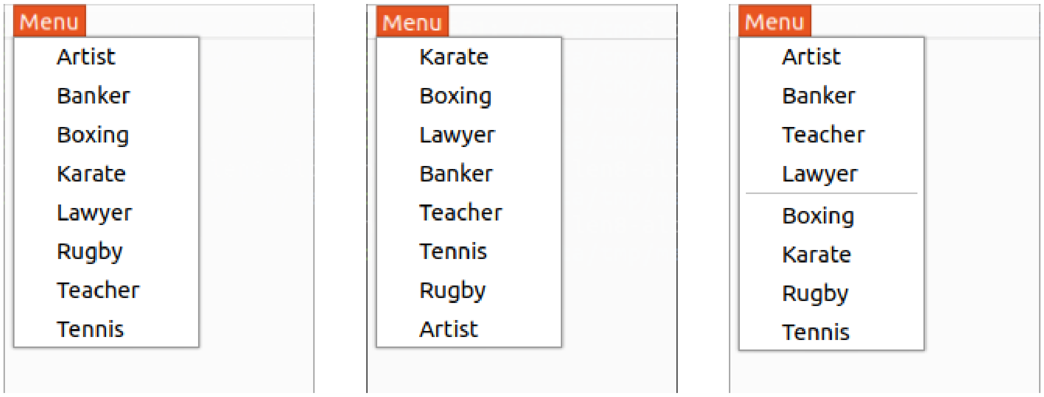


Figure D.11: Proportion of the three visual search strategies as a function of user id and menu organization for length 16.



(a) Alphabetic (b) Unordered (c) Semantic

Figure E.12: Examples of menu organizations (length 8, set B in Appendix E).