Journal of Experimental Psychology: General

ISSN: 0096-3445

2023, Vol. 152, No. 8, 2190–2204
https://doi.org/10.1037/xge0001390

How Do People Find Pairs?

Aoqi Li[1], Zhenzhong Chen[1, 2], Jeremy M. Wolfe[3, 4], and Christian N. L. Olivers[5, 6]
[1] School of Remote Sensing and Information Engineering, Wuhan University
[2] Hubei Luojia Laboratory, Wuhan, Hubei, PR China
[3] Brigham & Women's Hospital, Boston, Massachusetts
[4] Harvard Medical School, Harvard University
[5] Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam
[6] Institute for Brain & Behavior Amsterdam, Vrije Universiteit Amsterdam

Humans continuously scan their visual environment for relevant information. Such visual search behavior has typically been studied with tasks in which the search goal is constant and well-defined, requiring relatively little interplay between memory and orienting. Here we studied a situation in which the target is not known in advance, and instead, memory needs to be dynamically updated during the actual search. Observers compared two simultaneously presented arrays of objects for any matching pair of items—a task that requires continuous comparisons between what is seen now and what was seen a few moments ago. To manipulate the balance between memorizing and scanning, we ran two versions of the task. In an eye-tracking version, the objects were continuously available and could be scanned with relative ease. The results suggested that observers preferred scanning over memorizing. In a mouse-tracking version, perceptual availability was limited, and scanning was slowed. Now observers substantially increased their memory use. Thus, the results revealed a flexible and dynamic interplay between memory and perception. The findings aid in further bridging the research fields of attention and memory.

Public Significance Statement
Human observers can search their environment on the basis of abstract rules that demand a dynamic interplay between memory and perception, but this ability has received little investigation. This study uses the tools of the visual search literature to reveal how humans perform a task based on such an abstract rule; in this case, "find any matching pair of objects." The results demonstrate a continuous and flexible trade-off between internal memorizing and external perceptual sampling, depending on the balance of costs associated with these processes.

Keywords: pair search, short-term memory, long-term memory, trade-off

Supplemental materials: https://doi.org/10.1037/xge0001390.supp

To be as adaptive as possible, the visual system selects information that is relevant to the goals and needs of the organism, at the expense of irrelevant information. Such selection mechanisms have been extensively studied using visual search tasks, in which observers are asked to look for a target object in an array of distractor objects—akin to finding a spatula in the kitchen drawer, or a familiar face in a crowd. By the nature of the task, visual search requires a continuous interplay between memory (for what one is looking for) and perception (of what one is currently presented with). Standard theories of visual search therefore assume the activation of a representation of the search target in either working memory or long-term memory, which then guides or biases attention toward

This article was published Online First March 23, 2023.
Aoqi Li https://orcid.org/0000-0001-6373-6638
Part of the data in this article was previously presented in a poster at the annual meeting of the Vision Sciences Society. The preprint of this article is posted on PsyArXiv (preprint doi: https://psyarxiv.com/568q9/). All data have been made publicly available via OSF and can be accessed at https://osf.io/bvnjg/.

Aoqi Li was funded by the China Scholarship Council. Zhenzhong Chen was funded by the National Natural Science Foundation of China (Grant 62036005). Jeremy M. Wolfe was funded by the National Science Foundation (NSF; Grant 1848783) and the National Institutes of Health (Grant EY017001). Christian N. L. Olivers was funded by the Dutch Research Council (NWO; Grant 453-16-002).

Christian N. L. Olivers developed the study concept. Aoqi Li, Jeremy M. Wolfe, and Christian N. L. Olivers contributed to the study design. Aoqi Li conducted testing and data collection. Aoqi Li analyzed and interpreted the data under the supervision of Christian N. L. Olivers and Jeremy M. Wolfe. Aoqi Li drafted the manuscript. Zhenzhong Chen, Jeremy M. Wolfe, and Christian N. L. Olivers provided critical revisions. All the authors approved the final manuscript for submission.

Correspondence concerning this article should be addressed to Aoqi Li, School of Remote Sensing and Information Engineering, Wuhan University, Luoyu Road 129, Wuhan, Hubei, PR China, 420079. Email: aqli@whu.edu.cn

memory-matching features in the sensory input (e.g., Carlisle et al., 2011; Chan & Hayward, 2013; Duncan & Humphreys, 1989; Olivers et al., 2011; Wolfe, 2021; Woodman & Chun, 2006; Yu, Hanks, & Geng, 2022).

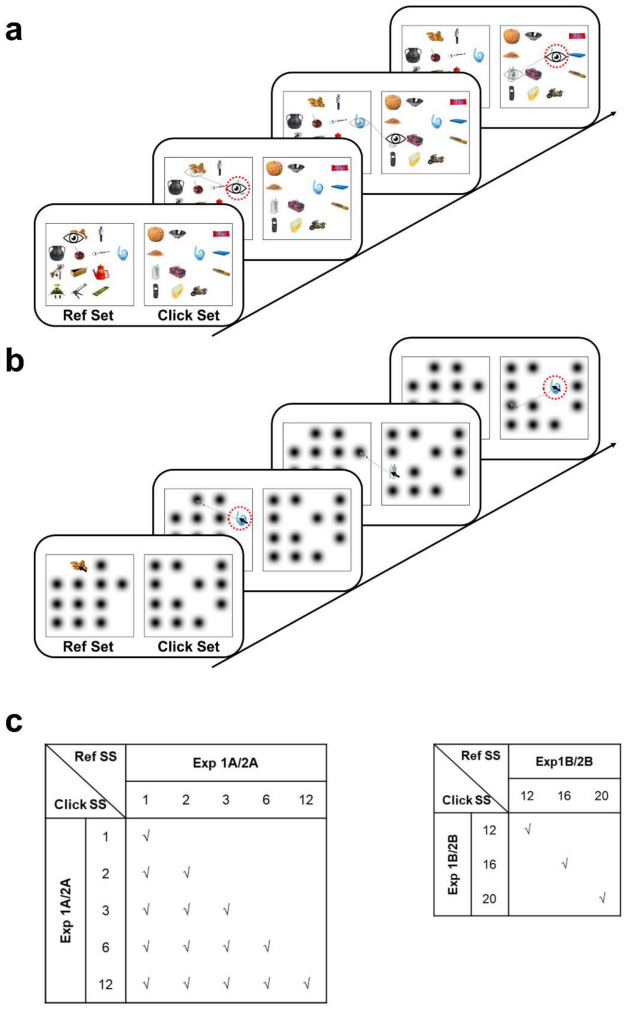
Visual search behavior has typically been studied under conditions where the object one is looking for is known in advance and the memory is therefore relatively fixed: The participant is given a specific target instruction (e.g., "search for the red square"), which then remains the same for at least the duration of the trial, but more often for a block of trials or even an entire experiment. It is probably fair to say that overall, researchers have taken the target memory as a given, and have focused on the visual selection process itself—thus ignoring the memory aspect of the search process (although there are exceptions, e.g., Alfandari et al., 2019; Carlisle et al., 2011; Wolfe, 2012). Yet, in natural behavior, what to look for may not always be so concretely defined (Alexander & Zelinsky, 2011; Bravo & Farid, 2009; Intraub, 1981; Maxfield & Zelinsky, 2012), and observers may adopt different representations for the same target, depending on the task and the other items in a display it needs to be distinguished from (Becker et al., 2010; Geng & Witkowski, 2019; Geng et al., 2017; Hout & Goldinger, 2015; Kerzel, 2019; Navalpakkam & Itti, 2007; Witkowski & Geng, 2019; Yu, Johal & Geng, 2022; see Yu et al., in press). In addition, there are real-life situations requiring a more flexible, dynamic use of memory during the search task itself. Imagine the scenario where one is looking for any two matching socks in a disorganized sock drawer. The observer holds the abstract search goal of finding a pair in memory, while there is no concrete a priori description of any specific sock. Instead, while looking through the jumble, the observer needs to continuously update their memory of what they have just seen in order to search for a potential match with what is looked at next. Little is known about how observers solve such tasks in which continuously changing perceptual input is compared to continuously changing memory, on the basis of an abstract overarching rule. To address this, in the current study, we investigated how humans search for pairs.

Figure 1a illustrates the core task. Observers were asked to search for a single pair of matching items across two simultaneously presented arrays of objects. That is, there was one object in the right display that had a matching counterpart in the left display. The task was to find the match and click on the counterpart. How might observers solve this task? A priori there are a number of strategies. First, observers might first exhaustively scan through and memorize one set (e.g., the array on the left in Figure 1a) and then move over to search through the other set (i.e., the array on the right). To apply this strategy, especially as set sizes get larger, one would have to make use of a large capacity, relatively persistent memory system. Work on recognition memory for objects makes it clear that a long-term memory exists and can rapidly encode many objects (Brady et al., 2008; Standing, 1973). Previous work has shown that this large capacity memory can be recruited in serving visual search (Wolfe, 2012). In what has become known as "hybrid search," observers can memorize dozens of candidate targets and then search from all of them at the same time within arrays of distractor objects. If this strategy is indeed being adopted in pair search, we expect extensive initial scans of the reference display and few transitions between displays.

Alternatively, observers could choose to limit memory usage. In the most extreme version of such a model (cf. Olivers et al., 2011), an observer would select one object from the left set, then search through the entire right set. If the target is not found, they would go back to the left set to select and remember the next potential target, repeating this process until the match is found. Thus, this strategy predicts limited scanning of the reference display, combined with many transitions between displays. Finally, memory usage may be adapted to a level between these extremes. For example, current memory usage could be limited to visual working memory capacity, which is typically around three to four objects (Cowan, 2001; Luck & Vogel, 2013).

Importantly, the memory strategy that observers choose may depend on multiple factors that are associated with the various task components, specifically the costs of associated actions

Figure 1
Example Trials and Experiment Design



Note. (a) Graphic illustration of a trial in the eye-tracking version of the experiments. The trajectory of the eye icon represents eye movements. (b) Graphic illustration of the mouse-tracking experiments. The trajectory of the mouse arrow represents mouse movements. Red circles (in both versions) indicate that the target object is selected. (c) All possible combinations of reference set size and click set size. See the online article for the color version of this figure.

(Hayhoe, 2017). Several studies on working memory have delved into this topic by use of an object-copying task. For example, some researchers varied the extent of movement required to accomplish the task (Ballard et al., 1995; Draschkow et al., 2021) while others varied the delay to stimulus availability (Somai et al., 2020). Each of these factors affected the use of memory, indicating that behavior is determined by a trade-off between cognitive effort and action-associated physical or time costs. To explore to what extent such principles can apply to visual search, we ran two versions of the pair search task, which are illustrated in Figure 1a and 1b, and which differed in two task components that might influence the memory strategy: (a) the perceptual availability of objects (which determines the amount of visual information that can be used to guide scanning); (b) the motor costs of eye/mouse movements. In the eye-tracking versions of the experiment (Experiments 1a and 1b), all objects in both arrays were visually available from the start of the trial. The continuous perceptual availability of the objects and the low motor costs of eye movements may foster a strategy that minimizes memory use, while maximizing scanning, as the world could be used as an external memory (cf. O'Regan, 1992). In contrast, in the mouse-tracking versions of the experiment (Experiments 2a and 2b), all objects were continuously covered from view, and only one object was revealed at a time, when the mouse pointer was hovering above it. The covering of objects was meant to minimize the visual information available for attentional guidance and, in combination with the mouse movements this made consulting the outside world more expensive. In this case, therefore, we expected to see a stronger reliance on (internal) memory.

To compare the overall search performance in both types of task, reaction times (RTs) were analyzed as a function of set size (i.e., "search slopes"). To assess the trade-off between external sampling and internal memorizing, we took the number of transitions between the two arrays and the total dwell time within a display before the observer transitioned to the other display as the main measures. The first measure is a proxy for external sampling, that is, the number of times the observer consulted the external world for information. The more information that is remembered from a display, the fewer transitions between displays are necessary. The second measure serves as the indication of memory encoding—under the assumption that the longer observers spend in a display, the more they try to remember. Finally, we used simulation models to estimate the number of memory observers used in this task and whether this depends on task-specific strategies.

Method

Participants

Two eye-tracking experiments (1a and 1b) and two mouse-tracking experiments (2a and 2b) were conducted. For Experiment 1a, we tested 25 participants (eight male, 17 female, $M_{age} = 22$, $SD = 3.65$, min = 19, max = 35), among whom one participant was added as a replacement for a participant who failed the calibration. For Experiment 1b, we tested another 25 participants (five male, 20 female, $M_{age} = 22.5$, $SD = 3.39$, min = 18, max = 29). All participants had normal or corrected-to-normal vision and were naïve to the purpose of the experiments.

Due to the COVID-19 pandemic, our laboratories were closed, and the two mouse-tracking experiments were therefore run online. Considering data from the online experiments might be noisier than data from the standard laboratory experiments, we increased the sample size to 35. For Experiment 2a, we tested 35 participants (four male, 29 female, two undetermined, $M = 20$, $SD = 1.78$, min = 18, max = 24). For Experiment 2b, we tested another 35 participants (four male, 31 female, $M = 20$, $SD = 2.59$, min = 18, max = 31).

All participants signed an informed consent form before they began the experiment. Participants received monetary compensation or credits for participation afterwards. The protocol was approved by the Scientific and Ethics Review Board of the Faculty of Behavioral and Movement Sciences of the Vrije Universiteit Amsterdam.

Stimuli and Apparatus

The eye-tracking experiments (Experiments 1a and 1b) were programmed in Matlab with Psychtoolbox-3. As shown in Figure 1a, the stimuli, consisting of two object sets in two side-by-side boxes, were displayed on a 24-in AUS XG248Q monitor with a refresh rate of 239 Hz and a spatial resolution of 1,920 × 1,080 pixels. For eye-tracking Experiment 1a, each box was partitioned into an invisible 4 × 4 grid and each object fit in a square of 96 × 96 pixels centered within a grid of 192 × 192 pixels. For eye-tracking Experiment 1b, each box was partitioned into an invisible 5 × 5 grid and each object fit in a square of 85 × 85 pixels centered within a grid of 154 × 154 pixels. In each trial, objects were randomly chosen from a public dataset of 4,760 objects (Konkle et al., 2010). Each object only appeared once (whether as a target object or as a distractor) during the experiment. Participants sat with their heads immobilized in a chin rest at a distance of about 70 cm from the screen. One degree of visual angle corresponded to about 44 pixels on the screen. Eye movements were recorded using an EyeLink 1000 Plus system (SR Research Ltd, Ontario, Canada) with a sampling rate of 1,000 Hz. A 9-point calibration and validation was done at the beginning of each block. Three participants in Experiment 1a and one participant in Experiment 1b failed the 9-point calibration. For these participants, 5-point calibration was conducted instead.

The online mouse-tracking experiments (Experiments 2a and 2b) mimicked the eye-tracking experiments (Experiments 1a and 1b), except as noted below. The mouse-tracking experiments were programmed in javascript with the jsPsych library and were run on the online platform, Cognition. The stimuli consisted of two object arrays presented in two side-by-side boxes. For Experiment 2a, each box was composed of a 4 × 4 grid and each object fit in a square of 81 × 81 pixels within a grid of 90 × 90 pixels. For Experiment 2b, each box was composed of a 5 × 5 grid and each object fit in a square of 72 × 72 pixels within a grid of 80 × 80 pixels. In each trial, objects were chosen from a public dataset of 2,400 objects (Brady et al., 2008). Due to a large number of trials and the limited number of pictures in this database, in Experiment 2a, each object only appeared once as a target but could return as a distractor. In Experiment 2b, each object only appeared once.

Design and Procedure

As shown in Figure 1a, the task involves searching for a single matched pair from the two arrays of objects. In both the eye-tracking (Experiments 1a and 1b) and the mouse-tracking experiments (Experiments 2a and 2b), participants were instructed to start from

the left display (reference display) and click on the paired target in the right display (click display). For Experiments 1a and 2a, set sizes of each object displayed varied independently. Set size could be 1, 2, 3, 6, or 12 with the constraint that the set size of the left-hand display was always equal to or smaller than the right-hand set size. For Experiments 1b and 2b, set sizes of both object displays were 12 on each side, 16 or 20. All possible combinations of set sizes are presented in Figure 1c.

In the eye-tracking experiments (Experiments 1a and 1b), all the objects were masked by Gaussian blobs at the beginning of each trial. Participants were asked to hover the mouse on any object in the left display to start the trial. Once a trial was started, all the objects became visible, as shown in Figure 1a. Participants were instructed to click on the target object in the right display when they found the match. Any click on a nontarget object would also end the trial and incur a time penalty. For Experiments 1a and 1b, the time penalty would increase 2 s with each additional error until it reached an upper limit of 11 s. The time penalty was reset at the beginning of each block. For Experiment 1a, there were 15 conditions differing in the set sizes of the two object displays. Each condition consisted of 35 target-present trials. All 525 trials were randomly intermixed across five blocks of 105 trials. For Experiment 1b, there were three conditions with 40 target-present trials per condition. All 120 trials were mixed across three blocks of 40 trials.

In the mouse-tracking experiments (Experiments 2a and 2b), all the objects were masked by Gaussian blobs during the whole trial. Participants were asked to move the mouse onto any object in the left display to start the trial. Once a trial was started, only one object would appear at a time when the mouse was hovering above the Gaussian blob placed at the corresponding position, as shown in Figure 1b. Participants were instructed to click on the target object in the right display when they found the match to an item from the left. Any click on a nontarget object would also end the trial and incur a time penalty. For Experiment 2a, the time penalty would increase 2 s with every error. For Experiment 2b, the time penalty would increase 3 s with every error but would be reset at the beginning of each block. For Experiment 2a, there were 15 conditions differing in the set sizes of the two object displays. Each condition contained 22 target-present trials. All 330 trials were mixed across five blocks of 66 trials. For Experiment 2b, there were three conditions with 20 target-present trials per condition. All 60 trials were mixed across six blocks of 10 trials.

Data Processing

Data Exclusions

For one participant of Experiment 1a, the calibration was substantially off in one block, and that block was removed from the dataset (0.8% of the total data). In Experiment 2a, about 0.7% of trials were removed due to a technical problem with the online testing.

In Experiments 1a and 1b, accuracy was generally high, and no participants were excluded (see Appendix Figure A1 in the online supplemental materials). As shown in Appendix Figure A2 in the online supplemental materials, two participants in Experiment 2a and two participants in Experiment 2b made more than 50% errors in at least one of the conditions, and their data were excluded to prevent a disproportionate speed–accuracy trade-off. At the trial level, in both experiments, for each participant, trials with incorrect responses were excluded from the analysis. We also excluded RTs that were ± 2.5 *SD*s from the mean for each condition. With rules, 96.3% and 94.7% of the total data remained for Experiment 1a and Experiment 1b, respectively, and 88.4% and 84.3% of the total data remained for Experiment 2a and Experiment 2b, respectively. Finally, note that to simplify analyses, participants were instructed to start their scan in the left display and click on the target in the right display. However, in the eye movement version, there were still a small number of trials in which the first fixation landed on the right display. To simplify the eye movement analyses, we excluded these trials for eye movement-related analyses (while retaining them for the RT analyses). This amounted to 1.5% and 1.1% of the total data in Experiment 1a and Experiment 1b, respectively.

Dependent Measures

For both the eye-tracking and the mouse-tracking experiments, we focused on three primary dependent measures: manual RT (defined as the time it took participants to find the right member of the pair and click on it) and its slope across set sizes as a measure of search efficiency, the number of transitions between displays (related to the number of times participants returned to a display to sample external information) as a proxy for external sampling, and the total dwell time before each transition (related to the number of objects participants successfully encoded into internal memory) as a proxy for internal memorizing. The latter was expected to correlate with the number of mouse/eye movements, which we also analyzed as secondary measures in Appendix B in the online supplemental materials.

Response Times

Search efficiency was analyzed using slope values derived from linear regression models fit on the RTs, as a function of reference and click set sizes. Slope values were then compared against zero using one-tailed one-sample *t* tests. Slope values for the reference and click set sizes were compared to each other using two-tailed paired *t* tests. Performance in the two different types of experiment (eye tracking vs. mouse tracking) was then also directly compared using two-way mixed ANOVAs with set size as the within-participant factor and experiment (Experiments 1 and 2) as the between-participant factor. Because in Experiments 1a and 2a, the reference set sizes and click set sizes were not manipulated orthogonally, we combined them into a single factor reflecting the combination of two set sizes.

Eye and Mouse Movement Analyses

For eye-tracking experiments, gaze data were parsed into saccades and fixations by the EyeLink online parser. Velocity and acceleration thresholds were set as 35 (°/s) and 9,500 (°/s^2), respectively. For the mouse-tracking experiments, only the sequences of visited objects and visit durations (how long each object was revealed by the mouse) were registered.

Transitions refer to the eye movements or mouse movements from the left-hand reference display to the right-hand click display or vice versa. In the eye-tracking experiments, a transition was defined as eye position crossing the midline from the left to the right or vice versa. The number of transitions was counted by the frequency of

saccades crossing the midline. In the mouse-tracking experiments, a transition was defined as the mouse moving from a left object to a right object or vice versa.

The dwell time before each transition was the cumulative amount of time during which participants inspected the objects in one display before transitioning to the other. In the eye-tracking experiments, since participants could get little information during fast saccadic eye movements, the dwell time before each transition was defined as the sum of all fixation durations in one display before a transition to the other display. In the mouse-tracking experiments, the dwell time before each transition was calculated as the sum of time during which participant hovered the mouse above a masked location to reveal the hidden object in one display before a transition to the other.

Same as for the RTs, both the number of transitions and the dwell times were then analyzed using linear regression models, of which the slopes were compared using one-sample t tests and paired t tests. The eye and mouse experiments were compared using a mixed ANOVA.

Note that on a small number of trials in the eye movement version, observers ended with the last fixation on the left display, even though they successfully clicked on the target on the right display. How these trials were dealt with depended on the specific situation. When the trial contained only fixations on the left, we assumed that participants already clicked before their eyes had landed on the target, and therefore we simply added one fixation on the right. We did the same under the scenario where observers had visited the right display just before ending on the left display, but they failed to land close to the target (i.e., within the arbitrary criterion of 3 dva). In contrast, if one of the fixations in the right display was close to the target (i.e., within the same arbitrary criterion of 3 dva), we counted that fixation on the right as the target fixation and disregarded all left display fixations at the end of the trial.

Transparency and Openness

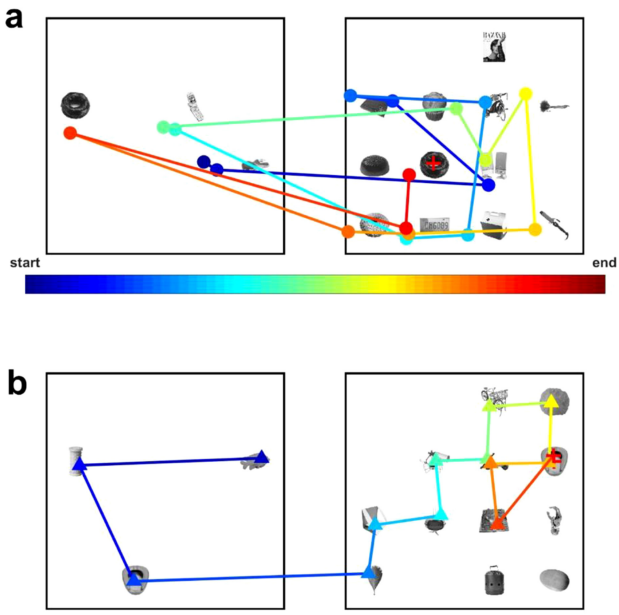
The design and analysis plans for the experiments were not preregistered. All data have been made publicly available via OSF and can be accessed at https://osf.io/bvnjg/.

Results and Discussion

Figure 2a and 2bb illustrate two typical search trajectories, as taken from example trials from Experiment 1a (eye tracking) and Experiment 2a (mouse tracking). While the combination of set sizes (3, 12) was the same for both trials, the search behavior is rather different. In Figure 2a (eye experiment), the observer chose to search in three cycles, each time picking one object from the reference display and then looking for it in the click display before returning to the reference display to select the next object. This pattern would be consistent with a strategy characterized by limited memory use, at the expense of more sampling. In contrast, in Figure 2b (mouse experiment), the observer completed the task in one cycle, as s/he chose to first look at all three objects in the reference display and then make only a single transition to the search display. This pattern is more consistent with a strategy that reduces the sampling, and instead makes more use of memory.[1]

To assess the search behavior in this type of matching task in general, and whether such behavior differed systematically for the two scanning modalities (eye vs. mouse), we focused on three primary dependent measures: (a) manual RT, including its slope across set sizes as an overall measure of search efficiency; (b) the number of transitions between displays as a proxy for external sampling; and (c) the dwell time before each transition, as a proxy for attempted memory encoding.

Figure 2
Visualization of Search Trajectories



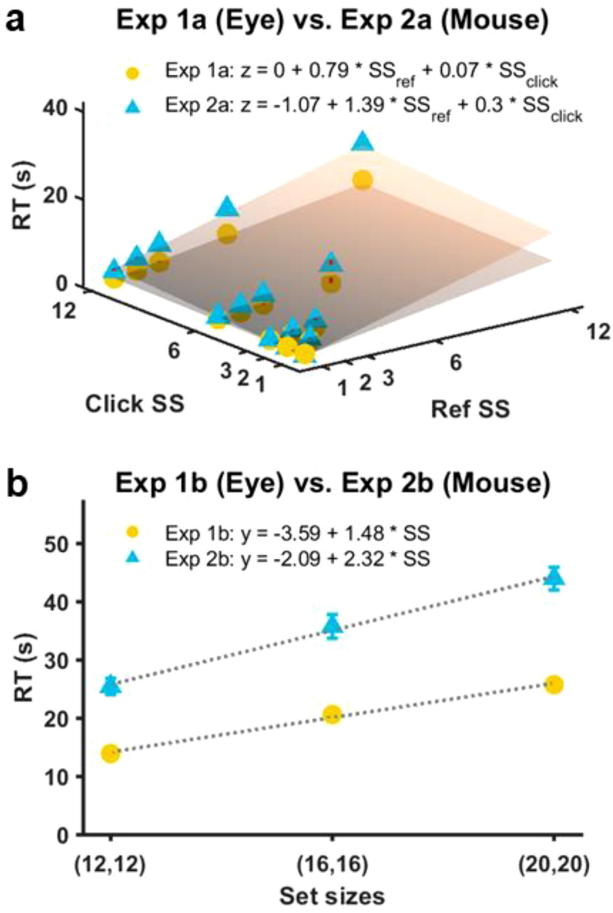
Note. (a) Example trial from Experiment 1a (eye tracking: Participant04-Block1-Trial40). (b) Example trial from Experiment 2a (mouse tracking: Participant01-Block1-Trial54). Object images are grayscale here to facilitate visualization (they were full color in the experiments). The colors of markers (dots for the eye-tracking trial and triangles for the mouse-tracking trial) and lines code the order of fixation, moving through blue–green–yellow–red. The final red cross indicates the click position. Note that the set size combinations are the same for both trials (3, 12), but the search behavior is quite different. See the online article for the color version of this figure.

Reaction Time

Figure 3a shows the group average of RTs plotted as a function of both reference and click set sizes for Experiment 1a (eye tracking) and Experiment 2a (mouse tracking). Figure 3b shows the group average of RTs plotted as a function of overall set size for Experiment 1b (eye tracking) and Experiment 2b (mouse tracking). We analyzed search efficiency by linear regression models of RTs as a function of the reference and click set sizes, for each individual participant, using the *fitlm* function in Matlab (min $R^2 = .88$, max $R^2 = .99$). In Experiment 1a, the slopes for the reference set size (795 ms/item) and the click set size (70 ms/item) were both significantly greater than zero as confirmed by one-sample t tests, one-tailed, reference display, $t(24) = 18.74$, $p < .001$, Cohen's $d = 3.75$; click display, $t(24) = 14.61$, $p < .001$, Cohen's $d = 2.92$. More

[1] Note that exhaustive scanning of a display does not necessarily mean exhaustive memorizing of that display. In a later section, we will use simulation models to estimate how much is being memorized from a scan.

Figure 3
Reaction Time



Note. (a) Reaction time as a function of set sizes in Experiments 1a (Eye) and 2a (Mouse). Red vertical lines represent the distance from data points to the fitted surface. (b) Comparison between reaction times (RTs) in Experiments 1b (Eye) and 2b (Mouse). Error bars represent between-subject standard errors. The equations describe the intercept and the set size slopes for the best linear fits. See the online article for the color version of this figure.

interestingly, the slopes for the reference set size were significantly larger than those for the click set size (paired t test, two-tailed), $t(24) = 17.72$, $p < .001$, Cohen's $d = 3.54$. For Experiment 1b, the linear regression model (min $R^2 = .62$, max $R^2 = 1.00$) revealed a mean search slope of 1,478 ms/item, which was significantly different from zero (one-sample t test, one-tailed), $t(24) = 14.84$, $p < .001$, Cohen's $d = 2.97$.

For Experiment 2a, too, linear regression models fit to both set sizes were run for each participant (min $R^2 = .91$, max $R^2 = 1$). Here too the slopes for the reference set size (1,388 ms/item) and the click set size (298 ms/item) were both significantly greater than zero by one-sample t tests, one-tailed, ref, $t(32) = 20.42$, $p < .001$, Cohen's $d = 3.55$; click, $t(32) = 26.54$, $p < .001$, Cohen's $d = 4.62$, and the slopes for the reference set size were again significantly larger than those for the click set size (paired t test, two-tailed), $t(32) = 16.31$, $p < .001$, Cohen's $d = 2.84$. For Experiment 2b, the linear regression method (min $R^2 = .18$, max $R^2 = 1.00$) yielded a mean slope of 2,320 ms/item, which was significantly greater than zero (one-sample t test, one-tailed), $t(32) = 11.52$, $p < .001$, Cohen's $d = 2.00$.

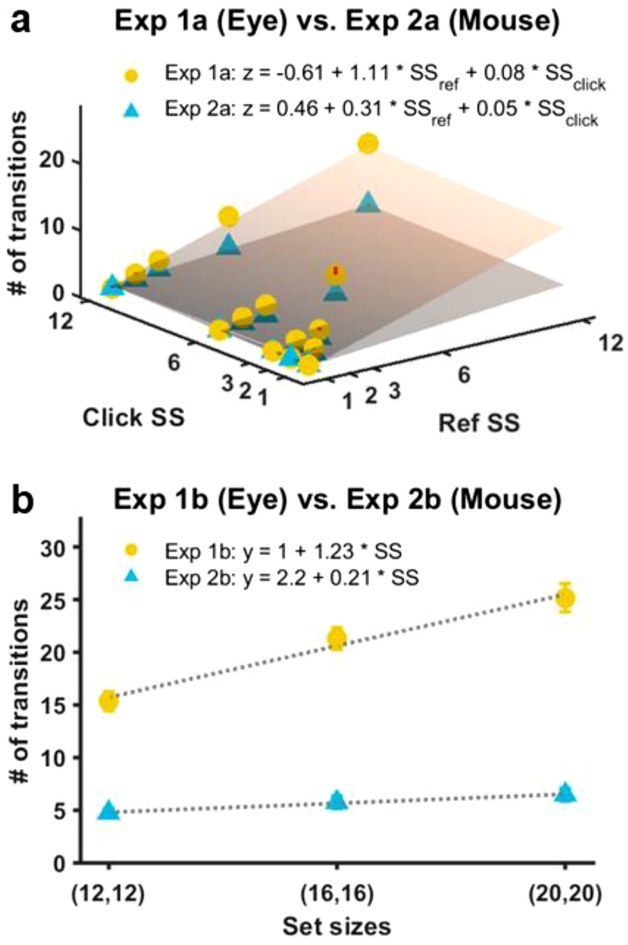
Finally, an ANOVA directly comparing the two modalities (eye tracking vs. mouse tracking) showed that in the mouse task, observers were overall slower (Experiment 1a vs. 2a, $F[1, 56] = 165.48$, $p < .001$, $\eta_p^2 = 0.75$; Experiment 1b vs. 2b, $F[1, 56] = 61.28$, $p < .001$, $\eta_p^2 = 0.52$) and less efficient (Experiment 1a vs. 2a, $F[14, 784] = 56.43$, $p < .001$, $\eta_p^2 = 0.50$; Experiment 1b vs. 2b, $F[2, 112] = 5.79$, $p < .01$, $\eta_p^2 = 0.09$).

A number of conclusions can be drawn. First, all the set size effects were linear in both versions of the task. This may have some implications for the way observers used their memory in this task, and we will return to this aspect in the General Discussion. Second, when the reference set size and the click set size were manipulated separately, RTs were much more affected by the reference set size than by the click set size. Note that this does not mean that observers actually spent ~800–1,300 ms looking at each object in the reference display. Rather, the reference set size acts as a multiplier: The more objects there are in the reference display, the more often observers also need to search through the click display. In any case, the differential slopes for the reference and click displays are consistent with observers largely treating them as the respective source and target for the comparison. Third, the mouse version of the task appears to be more effortful overall, as reflected in overall RTs as well as efficiency (slopes). This is not surprising given that (a) the mouse version involved objects that were covered until visited by the mouse, and (b) mouse movements themselves tend to be slower (and probably more effortful) than eye movements. The results thus confirm that the two versions involve different levels of effort in scanning the individual objects, which is important when we consider the exact scanning behavior and what it tells us about memory usage. We now turn to this scanning behavior.

Eye/Mouse Movements: Number of Transitions

Figure 4a shows the group average number of transitions plotted as a function of reference and click set sizes for Experiment 1a (eye tracking) and Experiment 2a (mouse tracking). Figure 4b shows the group average number of transitions plotted as a function of overall set sizes for Experiment 1b (eye tracking) and Experiment 2b (mouse tracking). For Experiment 1a, linear regression models of number of transitions as a function of both set sizes (min $R^2 = .86$, max $R^2 = .99$) showed that the slopes for the reference set size and the click set size were both significantly greater than zero, as confirmed by one-sample t tests, one-tailed, reference display, $t(24) = 22.97$, $p < .001$, Cohen's $d = 4.59$; click display, $t(24) = 11.59$, $p < .001$, Cohen's $d = 2.32$. The slopes for the reference set size were also larger than those for the click set size (paired t test, two-tailed), $t(24) = 20.98$, $p < .001$, Cohen's $d = 4.20$. For Experiment 1b, the linear regression model (min $R^2 = .38$, max $R^2 = 1$) revealed a mean slope significantly different from zero (one-sample t test, one-tailed), $t(24) = 10.59$, $p < .001$, Cohen's $d = 2.12$. The same fits for Experiment 2a (min $R^2 = .48$, max $R^2 = .98$) again revealed slopes for the reference set size and the click set size being significantly greater than zero by one-sample t tests, one-tailed, ref, $t(32) = 9.91$, $p < .001$, Cohen's $d = 1.73$; click, $t(32) = 9.40$, $p < .001$, Cohen's $d = 1.64$; plus the slopes for the reference set size were significantly larger than those for the click set size (paired t test, two-tailed), $t(32) = 8.52$, $p < .001$,

Figure 4
Number of Transitions



Note. (a) Transition number as a function of set sizes in Experiments 1a (Eye) and 2a (Mouse). Red vertical lines represent the distance from data points to the fitted surface. (b) Comparison between transition number in Experiments 1b (Eye) and 2b (Mouse). Error bars represent between-subject standard errors. The equations describe the intercept and the set size slopes for the best linear fits. See the online article for the color version of this figure.

Cohen's $d = 1.48$. For Experiment 2b, the linear regression (min $R^2 = .002$, max $R^2 = 1$) yielded a mean slope significantly greater than zero (one-sample t test, one-tailed), $t(32) = 5.11$, $p < .001$, Cohen's $d = 0.89$. Finally, an ANOVA directly comparing the eye-tracking and mouse-tracking experiments showed that in the latter type of task, observers made fewer transitions overall (Experiment 1a vs. 2a, $F[1, 56] = 226.76$, $p < .001$, $\eta_p^2 = 0.80$; Experiment 1b vs. 2b, $F[1, 56] = 209.36$, $p < .001$, $\eta_p^2 = 0.79$) and also fewer transitions per set size increment (Experiment 1a vs. 2a, $F[14, 784] = 141.56$, $p < .001$, $\eta_p^2 = 0.72$; Experiment 1b vs. 2b, $F[2, 112] = 49.65$, $p < .001$, $\eta_p^2 = 0.47$).

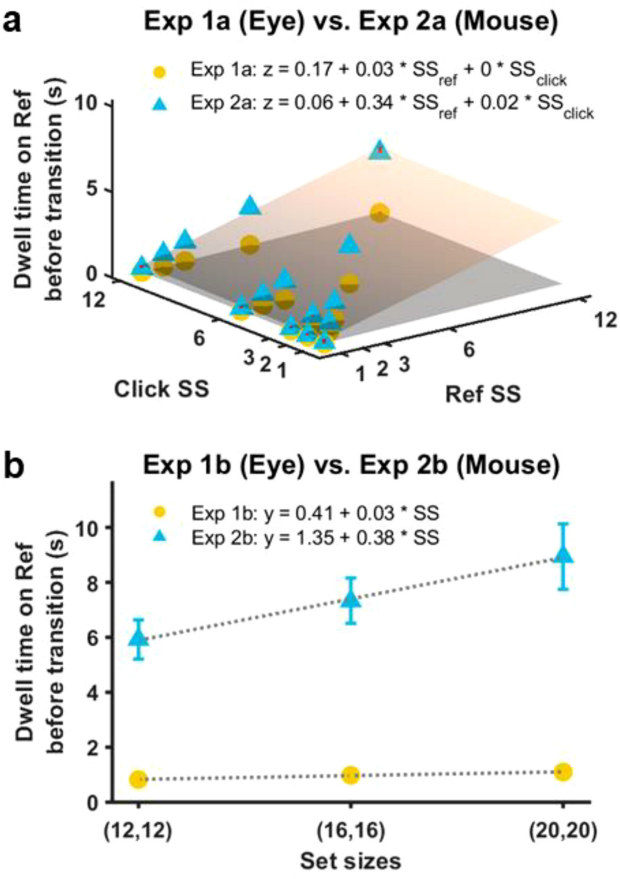
Thus, as for the manual RTs, here too we observe linear effects for both reference and click set sizes, and a stronger influence of the reference set size indicative that observers indeed used it as the reference. Importantly, the number of transitions was clearly reduced in the mouse version of the task.

Eye/Mouse Movements: Dwell Time Before Each Transition

For both eye-tracking and mouse-tracking experiments, we computed the average dwell time separately for the reference and the click displays. Here we focus on the reference display while dwell times for the click display are reported in Appendix Figure B1 in the online supplemental materials. Analyses on secondary measures underlying the dwell time, notably the number of visits/fixations and the average dwell time per visit/fixation are also shown in Appendix B in the online supplemental materials. Appendix Figures B2–B5 in the online supplemental materials show that the total dwell time was primarily driven by the number of fixations/mouse-visits, rather than the fixation/mouse-visit duration.

Figure 5a shows the group average of dwell time before each transition plotted as a function of reference and click set sizes for Experiment 1a (eye tracking) and Experiment 2a (mouse tracking).

Figure 5
Dwell Time Before Each Transition (ref)



Note. (a) Dwell time before each transition as a function of set sizes in Experiments 1a (Eye) and 2a (Mouse). Red vertical lines represent the distance from data points to the fitted surface. (b) Comparison between dwell time before each transition in Experiments 1b (Eye) and 2b (Mouse). Error bars represent between-subject standard errors. The equations describe the intercept and the set size slopes for the best linear fits. See the online article for the color version of this figure.

Figure 5b shows the group average of dwell time before each transition plotted as a function of overall set sizes for Experiment 1b (eye tracking) and Experiment 2b (mouse tracking). For Experiment 1a, linear regression models of dwell time as a function of both set sizes (min $R^2 = .86$, max $R^2 = .99$) showed that the slopes for the reference set size and the click set size were both significantly greater than zero as confirmed by one-sample t tests, one-tailed, reference display, $t(24) = 14.33$, $p < .001$, Cohen's $d = 2.87$; click display, $t(24) = 4.83$, $p < .001$, Cohen's $d = 0.97$. The slopes for the reference set size were also larger than those for the click set size (paired t test, two-tailed), $t(24) = 13.81$, $p < .001$, Cohen's $d = 2.76$. For Experiment 1b, the linear regression model (min $R^2 = .55$, max $R^2 = 1.00$) revealed a mean slope significantly different from zero (one-sample t test, one-tailed), $t(24) = 3.44$, $p < .01$, Cohen's $d = 0.69$. For Experiment 2a too, linear regression models fit to both set sizes were run for each participant (min $R^2 = .74$, max $R^2 = 1.00$). Here too the slopes for the reference set size and the click set size were both significantly greater than zero by one-sample t tests, one-tailed, ref, $t(32) = 13.14$, $p < .001$, Cohen's $d = 2.29$; click, $t(32) = 8.57$, $p < .001$, Cohen's $d = 1.49$; and the slopes for the reference set size were again significantly larger than those for the click set size (paired t test, two-tailed), $t(32) = 12.54$, $p < .001$, Cohen's $d = 2.18$. For Experiment 2b, the linear regression (min $R^2 = .0035$, max $R^2 = 1.00$) yielded a mean slope significantly greater than zero (one-sample t test, one-tailed), $t(32) = 5.31$, $p < .001$, Cohen's $d = 0.92$. Finally, an ANOVA directly comparing the eye-tracking and mouse-tracking experiments showed that in the latter type of task, observers dwelled overall longer before each transition (Experiment 1a vs. 2a, $F[1, 56] = 173.68$, $p < .001$, $\eta_p^2 = 0.76$; Experiment 1b vs. 2b, $F[1, 56] = 37.87$, $p < .001$, $\eta_p^2 = 0.40$), and dwell time increased more per set size (Experiment 1a vs. 2a, $F[14, 784] = 97.11$, $p < .001$, $\eta_p^2 = 0.63$; Experiment 1b vs. 2b, $F[2, 112] = 14.65$, $p < .001$, $\eta_p^2 = 0.21$).

Thus, as was the case for the other dependent measures, for dwell time too we observe linear effects for both reference and click set sizes, and a stronger influence of the reference set size. Importantly, while the number of transitions decreased for the mouse experiment compared with the eye experiment, the dwell time increased. This is suggestive of a trade-off between dwelling and transitioning across tasks.

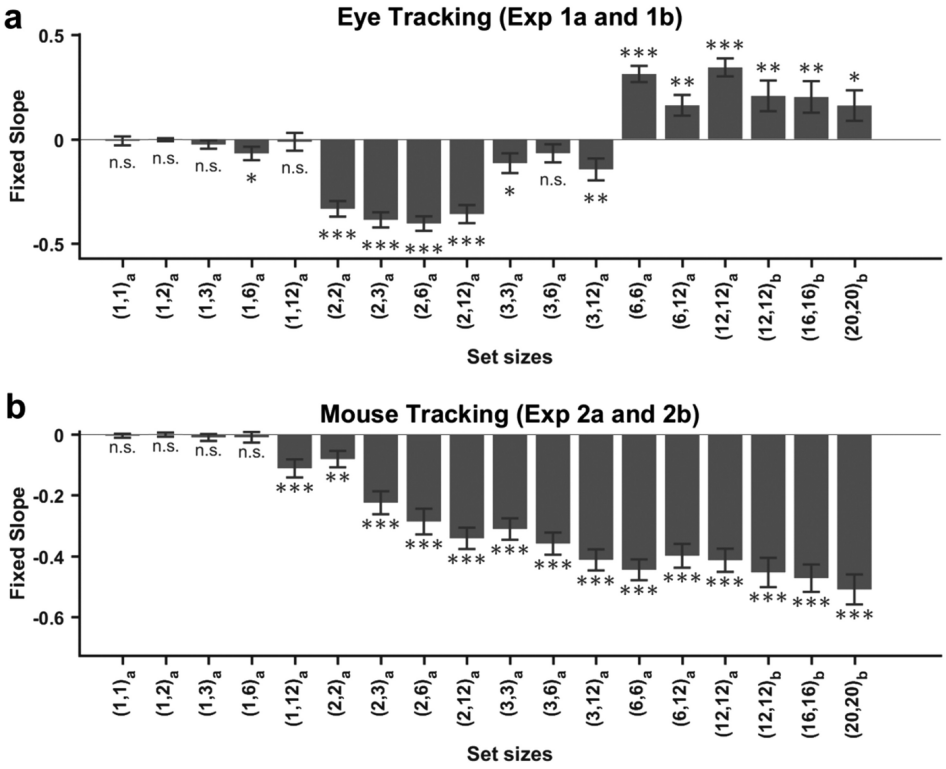
To find further support for such trade-offs, we also assessed the relationship between dwelling and transitioning behavior *within* each of the experiments. For this purpose, we applied a linear mixed effects (LME) model to the individual trial data for each of the set size combinations, to see if longer dwell times indeed predicted fewer transitions. The model structure we adopted was #transitions $\sim 1 +$ dwell_time $+ (1 +$ dwell_time|Participant), thus treating number of transitions as the dependent variable, dwell time as the predictor and participant as a random factor for which both intercept and slope were estimated. For each condition, data were z-scored within each participant before being fed into the model. Figure 6 shows the LME slope results, reflecting the relationship between dwell time and number of transitions for each of the set size combinations of Experiments 1 and 2. Negative slopes indicate a trade-off where longer dwell times indeed go with fewer transitions. Such trade-offs were not universally present for the eye-tracking task (Experiments 1a and 1b), as is shown in Figure 6a. At the smallest set sizes, there was no significant relationship between dwell time and number of transitions. The intermediate set sizes did show a negative slope, indicative of a trade-off. However, at the larger set sizes, the relationship turned positive indicating that a larger number of transitions also went together with longer dwell times. This pattern is somewhat difficult to interpret, and may mean that people changed tactics when reference set sizes got quite large. The pattern is clearer for the mouse task (Experiments 2a and 2b), as here virtually all slopes were negative (and reliably so from set size [1, 12] onwards) from Figure 6b. Thus, in the mouse task, there was a more pronounced trade-off between dwelling and transitioning consistent with changing sampling/memorizing strategies, in this case from trial to trial. Detailed statistics of the LME results are presented in Appendix Table C1 in the online supplemental materials.

An additional question we can then ask here is what is most beneficial for search: dwelling or transitioning? We therefore repeated the analyses but now, instead of the number of transitions, we tried to predict RTs from dwell time. The results are shown in detail in Appendix Table C2 in the online supplemental materials. Across the board, the relation between dwell time and RTs was positive, suggesting that investing more in memorizing is not necessarily the most effective way of completing the task.

Estimating Memory Usage Using Simulation Models

The observed differential trade-offs between dwelling and transitioning for the eye and mouse experiments indicate different memorizing strategies. Observers transition more between the two displays in the eye movement version of the experiment, suggesting that they rely more on the sensory input, and less on memory, while the reverse appears true for the mouse version. But how many objects did observers then actually store in memory when switching from one display to the other? The slope values of the number of transitions can already help us exclude two extreme search strategies, a strategy where people first fully learn the reference display by heart before they search through the click display, and the strategy where they limit memory to one object at a time. Under the first strategy, the total number of transitions would be constant regardless of the reference set size. This was clearly not the case. Under the second strategy, observers would visit only one object in reference display at a time before transitioning and search the other display for the single target, so the average number of transitions from the reference display would be the same as the expected number of visited objects in the reference display, which is $(SS_{ref} + 1)/2$, and the total number of transitions would then be $SS_{ref} + 1$ (given that observers also transition back from the click display). As shown in Figure 4, none of these predictions held, whether for the eye-tracking or the mouse-tracking version. To provide a more precise estimate of memory usage in the two versions of experiments, we built two simulation models which were highly similar, except for some task-specific adaptations. The core model is illustrated in Figure 7, while Appendix D in the online supplemental materials contains pseudocode. The model seeks to predict the number of transitions observers make, by varying the amount of information carried from one display to the next (i.e., memory usage). To this end, when the model scans one display, it randomly selects a subset of objects as visited (i.e., "seen") objects, referred to as n. From this subset of n visited objects, a further subset of objects, referred to as m, are then selected, which are being memorized across the transition to the other display. Thus, the model can only remember (a selection

Figure 6
Trade-off Between Dwelling and Transitioning



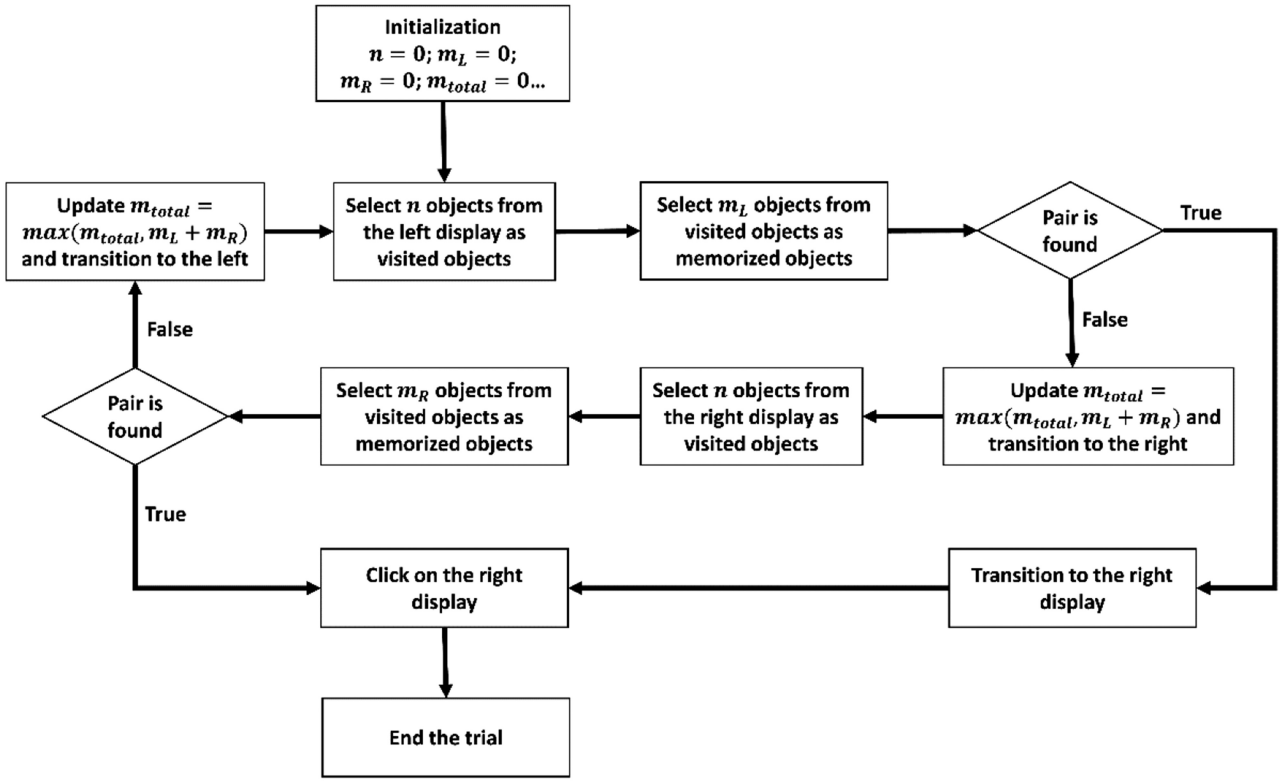
Note. (a) Data from Experiments 1a and 1b (eye tracking). (b) Data from Experiments 2a and 2b (mouse tracking).

of) what it has seen. The model then transitions to the other display, where it again selects a subset of visited objects. If both members of the pair are present in the selected subsets (one in memory, one in the currently seen set), the pair is found. If no pair is present, the model will update its memory with a subset of the visited objects in the currently viewed display, and then transition back to the other display. This process is repeated until the pair is found. Note the model is allowed to carry information in both directions, from the left (reference) display to the right (click) display and vice versa, from the right back to the left. As we tried—by design as well as instruction—to make participants treat the left display as a reference (or memory) display and the right display as the click (or search) display, we decided to model the amount of information being gathered from the left and right displays with two separate variables, referred to as m_L and m_R, where m stands for memory usage. How much from both displays will be remembered will dynamically depend from transition to transition on how much from the current and the previous displays is remembered, and will thus be a combination of m_L and m_R. We then take the maximum total number of objects that is being remembered on a particular trial as the estimate of the actual total memory usage that is being recruited for that trial, and we will refer to it as m_{total}.[2] Thus, for each simulated trial, memory usage is defined as the average number of memorized objects before each transition while the total memory usage is defined as the maximum number of objects ever simultaneously held in memory at any point during a trial.

The basic model makes a number of additional assumptions. First, it assumes no memory beyond one transition. That is, when the model transitions back to the left display, what the model selects as potential targets for subsequent scanning will not be influenced by what the model has seen earlier in a previous visit to that display, and the same is the case for the right display. This is commensurate with earlier work indicating that the effects of attention have no cumulative effect on visual perception (Horowitz & Wolfe, 1998, 2001; Wolfe et al., 2000). As shown in Appendices E1 and F1 in the online supplemental materials, the number of fixations or mouse-visits on the reference display before each transition decreased with the first one or two passings on the reference display, consistent with the possibility of some initial learning of the objects across transitions. However, beyond these initial rounds, the number of fixations/visits remained stable, suggesting little further memory build-up. There was also no sign of a learning effect for the click display. We therefore decided to keep our basis model simple, with no memory beyond one transition. For the mouse-tracking experiments, this simple model accounted for the data well. However, in the eye movement version, we deviated from this assumption, as will be

[2] Note that due to the model dynamics, the average m_{total} is not simply the sum of the average m_L and average m_R, as m_L and m_R are dynamically changing with each transition. Instead m_{total} is the maximum sum of m_L and m_R at one specific moment within a trial.

Figure 7
The Workflow of the Core Model



explained later. Second, when the model incidentally happens to remember *all* the objects from one display, it will transition to the other display and then always exhaustively search the other display for the match. This makes sense because when it remembers everything, there is no reason to return to the previous display and sample more. This rule especially helps mimic the human data for low reference set sizes, especially set size 1, when observers are highly likely to have perfect memory. The same rule has little impact for larger set sizes as the model is then unlikely to remember all the objects anyway.

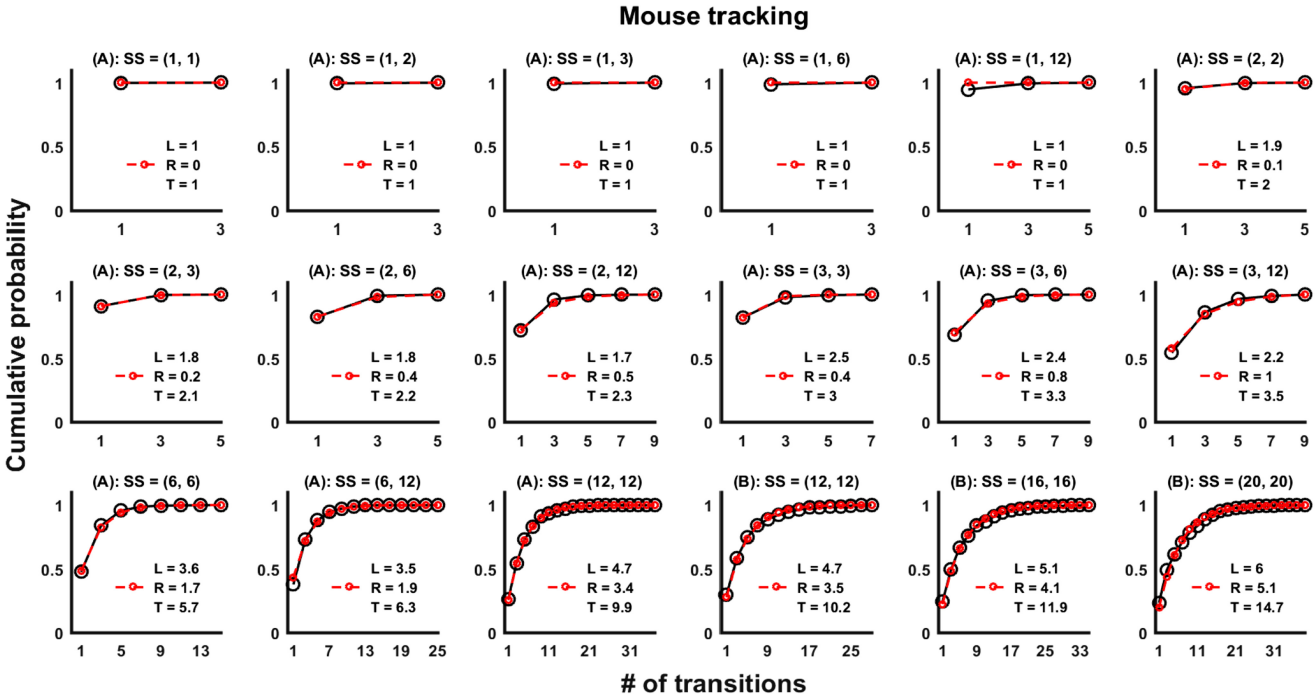
Modeling the Mouse-Tracking Data

We start with data from the mouse-tracking experiments, because for those experiments we have exact data on how many objects were visited in each display before each transition (variable n)—as each object would only become visible upon a mouse-visit. Appendix E2 in the online supplemental materials shows the distributions of the number of objects visited before each transition for the two displays in the mouse experiments. Note that the peak of the distribution generally appears at the corresponding set size, consistent with the idea that participants had a tendency to scan all the objects in one display before transitioning to the other (which is not the same as *memorizing* all the objects). The simulation model draws n directly and randomly from this distribution (i.e., the actual data on a number of visits). A subset from this draw, m, is then selected for memory. In the model fitting, the memory usage variables m_{L} and m_{R} were not varied directly though, as this would ignore random variation in capacity from moment to moment, and limit the precision of the model to linearly increasing integer steps. To allow for random fluctuations, as well as a nonlinear asymptotic increase in memory usage, on each transition, the memory usage for each sampling of a display was modeled as follows: $m_{\mathrm{L,R}} = \mathrm{MIN}(n,\ \mathrm{RAND}(1,\ u))$, where n is the number of objects seen, and u is a free parameter setting the upper memory boundary. Thus, memory usage was constrained by the number of objects seen, and a momentary memory limit as drawn from a uniform distribution with upper bound u, whichever of the two was smallest. Note that u is merely used to limit the memory capacity distribution here which indirectly affects m, and should not be interpreted as directly reflecting human memory capacity. This method of setting memory usage ensured a better fit than a simple linear function with integer steps. The only free parameter in the model is then u. The memory usage was thus set for each search cycle within a trial for the left (m_{L}) and the right (m_{R}) displays separately. The maximum memory usage within a trial was then m_{total}, which reflected the maximum combination of m_{L} and m_{R} during any cycle within the trial.

The model was then fitted to the cumulative distributions of transitions in the human data. We fitted 10 runs using a grid search method, with a thousand simulated trials for each run. Appendix E3 in the online supplemental materials shows the fit values for a range of parameter (u) settings. Figure 8 shows the average of best-fit models from 10 runs together with the human data, and the resulting estimates for m_{L}, m_{R}, and m_{total}. We will return to these memory values

Figure 8

The Average of Best-Fit Models From 10 Runs for Mouse-Tracking Data



Note. Black circles represent human data. Red circles represent simulated data. See the online article for the color version of this figure.

later. For now, we conclude that the model fits the data quite adequately.

Modeling the Eye-Tracking Data

To mimic transitioning behavior in the eye-tracking experiment, the model required a number of adaptations. The first adaptation concerned the number of objects "seen" by the model. Note that, except at initial onset of a display, the objects were not covered, and thus remained visible throughout the trial. As a consequence, while we can track and count individual fixations, we cannot unequivocally determine how many, and which, objects were "seen" within each fixation. Fixations were not always clearly directed to an object, as they could fall on empty space, but observers may also sample multiple objects within a single fixation, using extrafoveal vision. Because we could not retrieve from the data how many objects were really seen as such, we estimated this instead by adding a second parameter to the model. This parameter, *f*, reflects the radius of the functional viewing field, a circular area around fixation within which objects were regarded as "seen" (cf. Hulleman & Olivers, 2017; Wolfe, 2021; Young & Hulleman, 2013). It was allowed to vary between 3 and 15 dva. Thus, the larger *f*, the more objects could be seen per fixation.

The second adaptation concerned the introduction of an *inhibition of return* mechanism (Klein, 2000). Note that the basic model assumed no memory beyond one transition. This accounted well for the mouse data, where observers often chose to re-scan objects when they returned to a display. However, we observed a considerably weaker fit for the eye movement data. The reason is illustrated in the example trial shown in Figure 2. While in the eye experiment observers more often went back and forth between displays, they did not do so randomly. Rather, there appeared to be some memory of where one had already previously sampled. Especially at set size 2 (and to a lesser extent set size 3), observers realized that if it is not the first object they picked, it must be the other. This implies some spatial memory or strategy reflecting which objects can be safely ignored. To mimic this aspect, we implemented a rudimentary inhibition of return function which biased the model against memorizing previously memorized objects (see Appendix D in the online supplemental materials for details).

Appendix F2 in the online supplemental materials illustrates the mean squared errors of model fits from one run as a function of the now two free parameters *u* (the parameter from which *m* is derived), and *f* (the size of the functional viewing field; brighter colors indicate a better fit, i.e., least squared error). This also immediately illustrates another problem: Multiple parameter combinations generate relatively good fits, and different runs converged on somewhat different combinations. This happens because the effects of the parameters *u* and *f* are not independent: How much is actually being remembered (*m*) is determined by the minimum of how much can be stored at that moment (as determined by *u*), and how much was seen (as determined by *f*). After all, what is not seen can also not be remembered. Hence, the model may find a solution for a large *f* combined with a small *u*, the other way around, or something in between. Note though that the implication is the same: The amount of memory that is eventually being used during the task (i.e., *m*) is limited, either by the momentary memory capacity itself, or by the number of objects being encoded into that memory. We therefore solved this

problem by simply averaging the obtained parameter values across the 10 runs into a pooled estimate. Figure 9 shows the average of best-fit models from 10 runs to the cumulative distributions of transitions in the human data, together with the estimates of m_L, m_R, and m_{total}. Again, the fit appears very adequate. Appendix F3 in the online supplemental materials shows the best model fit when no inhibition of return is assumed for the eye data. While still capturing the overall pattern, this model provides a visibly worse fit than the model with inhibition of return for especially the reference set sizes 2 and 3. Conversely, Appendix E4 in the online supplemental materials shows the model fit for the mouse-tracking data, now with inhibition of return included. Here too the fit is worse, again mainly for set sizes 2 and 3. This provides another difference in memory strategy between the two tasks.

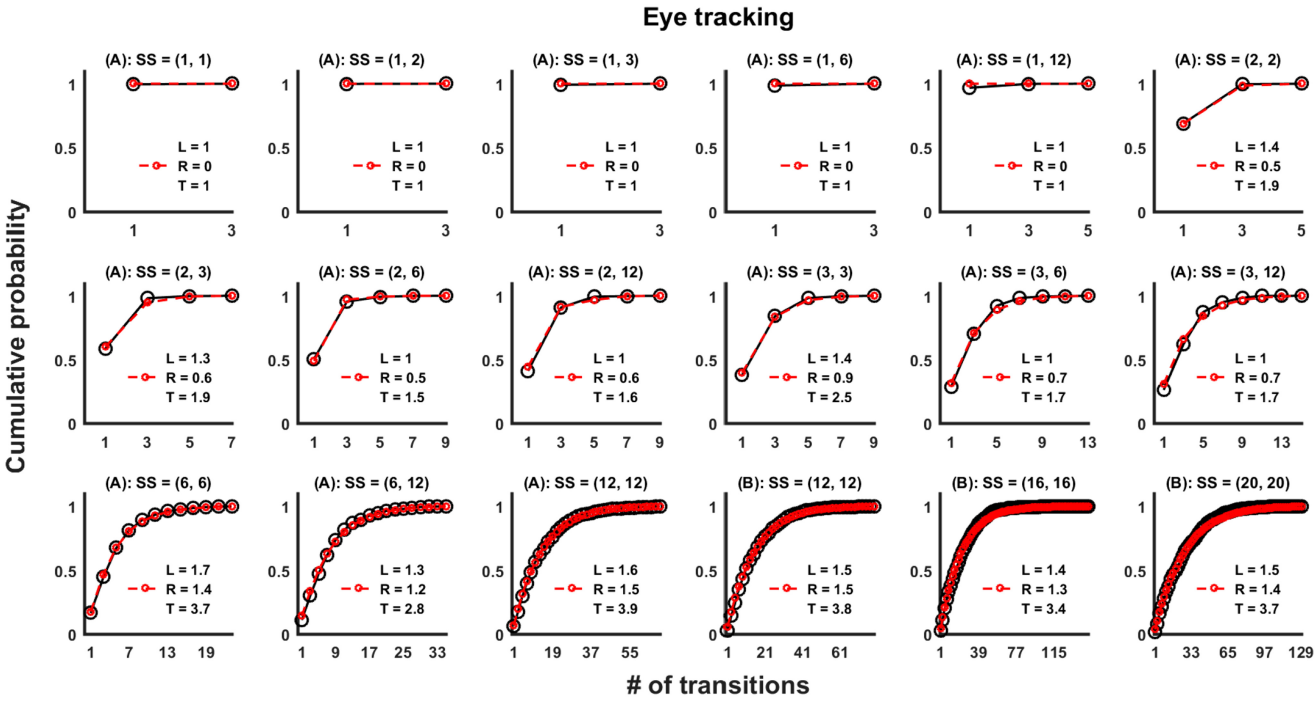
Comparing Memory Usage in the Eye- and Mouse-Tracking Experiments

Figure 10 shows the average estimated memory usage for both the eye-tracking and the mouse-tracking versions of the experiment, as a function of the reference set size and the click set size. Dashed lines and dotted lines represent estimated memory usage for the left display and the right display, respectively, whereas solid lines show the estimated total memory usage. A number of observations can be made. First, the number of objects remembered increases with the reference set size, with overall little to no additional effect of the click set size. This suggests again that observers indeed primarily treated the reference display as the memory display and the click display as the search display. Second, in the mouse-tracking version, overall more information was being remembered from the left display than from the right display, which would be consistent with a dominant "remember left, search right" strategy. However, in the eye-tracking version, there was no such difference, as similar amounts of information were carried back and forth between displays. Last but not least, in the eye-tracking experiments, total memory usage clearly plateaued with increasing set size, at a value close to four items. In contrast, in the mouse-tracking version, memory usage steadily grew with set size, with an average m_{total} of more than 14.5 for set size (20, 20; note that exact estimates may differ slightly per simulation due to the random components). And although the increase in memory usage with set sizes appears to slow down a little for higher set sizes, there is no asymptote in sight yet.

General Discussion

Compared with typical investigations of visual search tasks, where the target would usually be explicitly defined, our pair search task only defines the goal of finding a pair. The specific target on each trial is determined by the search stimuli themselves, rather than by instructions prior to search. In this type of task, observers cannot be guided by a fixed target representation in memory. They need to dynamically update representations of potential target candidates based on the visual input during scanning. We were interested in how observers coordinate scanning and memorizing in such a task. To address this question, we conducted two versions of the experiment (eye and mouse) in which observers searched in two arrays for a pair of matching objects, one member of the pair in

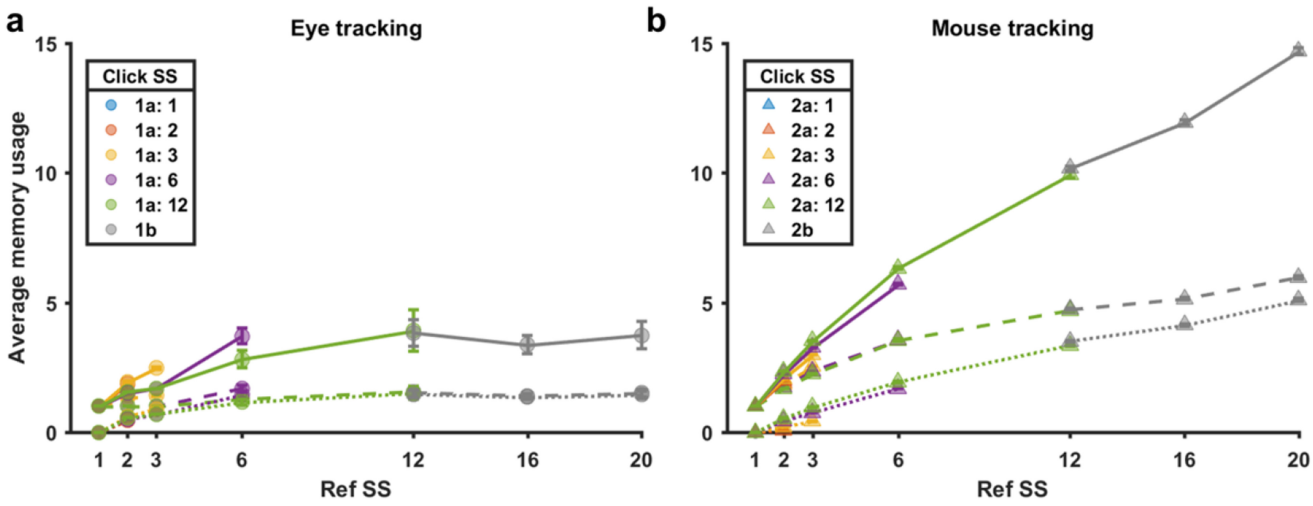
Figure 9
The Average of Best-Fit Models From 10 Runs for Eye-Tracking Data



Note. Black circles represent human data. Red circles represent simulated data. See the online article for the color version of this figure.

Figure 10

Estimated Memory Usage in (a) the Eye-Tracking Experiments and (b) the Mouse-Tracking Experiments, as Derived From the Model Simulations



Note. Estimated memory usage for the left display (dashed line) and the right display (dotted line) as well as estimated total memory usage (solid line) is shown, as a function of set sizes. See the online article for the color version of this figure.

each array. Between the two versions, we varied the costs that came with scanning the visual information, in order to assess if that resulted in different memory strategies.

When looking at task performance (as reflected in manual RTs), we observed all the set size effects were linear in both versions of the task, though RTs were much more affected by the reference set size than by the click set size when they were manipulated separately. The differential slopes for the reference and click sets suggested that observers largely treated the left display as the source of the "memory set" and the right display as the "visual set" as those terms are defined in hybrid search (Wolfe, 2012). Wolfe (2012) reported that RTs increased with the log of the memory set size. However, we did not observe a logarithmic relationship with the reference set size. This could suggest that the observers did not adopt a strategy of storing all items from the reference set in memory in the same way as they would perform a hybrid search. Alternatively, the experiment may not have been sufficiently sensitive to detect a log-linear memory search. Note that in the typical hybrid search task, observers are first required to commit all of the target items to memory before performing the visual search. Hybrid search RTs therefore do not include the time participants spent on memorizing objects. In our pair search task RTs do include the time of encoding the items in memory, and this may well be a strong linear effect, potentially masking any additional log-linear effects.

Of most interest was the number of transitions between displays and the dwell time spent in each display. First of all, these metrics indicate that observers indeed regularly updated their memory during the trial, as observers typically required multiple transitions between displays, especially for the larger reference set sizes. Second, these metrics suggest that observers approached the two types of tasks with very different strategies. Participants relied more on internal memorizing (fewer transitions and longer dwell times) in the mouse-tracking experiments but more on external sampling (more transitions and shorter dwell times) in the eye-tracking experiments, suggesting a strategic trade-off between sampling and memorizing across the two versions of experiments. This was further confirmed by direct estimates of memory usage, as derived from simulation models for each of the two tasks. The estimated total memory usage plateaued at about four in the eye version but grew to considerably higher values in the mouse version.

To summarize then, observers relied more on the outside stimuli when scanning was easy, and relied more on their internal memory when scanning was hard. The findings are consistent with the idea of the "world as external memory," as proposed by O'Regan (1992). As he pointed out, even if we have the memory capacity, we may not use it if the cost of simply reacquiring information from the external world is low enough. This theory has been supported by several earlier findings. For example, Ballard et al. (1995) used a block copying task in which participants were required to copy a model grid of blocks in another grid. It was found that participants sought to minimize the use of memory at the expense of making more eye movements between the model and the response grid, presumably because the cost of memory was more expensive than the cost of the eye movements. By further increasing the cost of acquiring information (the distance between the model and the workspace), the researchers found the strategy shifted in the direction of more memory use, showing that the cost of associated actions is an important factor determining how memory will be used (see also Draschkow et al., 2021). Such costs associated with additional actions may reflect direct motor costs, where memory usage is traded off against movements in terms of energy expenditure (Hayhoe & Matthis, 2018; Li et al., 2018), but may also reflect time costs (as larger or more protracted actions need more time). Evidence for a time component comes from a study that also used the block copying task, and which found that increasing the delay to stimulus availability also led to a shift from sampling (more eye movements from the model grid to the response grid) to memorizing (longer dwell time on the model

grid; Somai et al., 2020). Though it is different in a number of aspects, our current task bears a clear resemblance to the object-copying paradigm, so it is encouraging to see that the same principles also hold for visual pair search.

It is interesting to note that estimated memory usage stayed within a limit of four items in the eye-tracking experiments even when the reference set size was much >4. In contrast, memory use grew with set size to more than 14.5 items in the mouse version. As reported in other typical laboratory experiments about memory capacity, the capacity of short-term memory is limited to three or four objects (Cowan, 2001; Luck & Vogel, 2013) while the capacity of long-term memory can be massive (Brady et al., 2008; Standing, 1973). The simulated results about memory usage suggest that participants might have only relied on short-term memory in the eye-tracking experiments while engaging long-term memory in the mouse experiments. This conclusion is also supported by the differential memory usage for the left and right displays. In the eye-tracking version, the estimated usage was similar for both left and right sets, suggesting that participants were continuously modifying the contents of working memory. In the mouse version, the estimated usage for the left, reference display was larger overall than the usage for the right display, suggesting that participants were trying to adopt a strategy similar to that used in a hybrid search (here "remember left, search right").

Our model assumes no accumulation of memory across multiple scans of a display (i.e., it assumes no memory beyond one transition between displays). One may question this assumption given that previous studies have demonstrated that the identities of distractors can be incidentally remembered during visual search, as reflected in shorter search RTs and fewer fixations after search displays are repeated (Hout & Goldinger, 2010, 2012). We conducted a similar analysis at a more fine-grained level which looked at the number of scanning movements (eye or mouse) as a function of transition number (see Appendices E1 and F1 in the online supplemental materials), but our results only showed a decreasing trend for a limited number of cases, namely only for the reference display, only for the higher set sizes (12–20), and only for the first one (in the mouse version) or two (eye version) scans. Note further that scanning an item is not necessarily the same as remembering the item (as estimated by our memory usage parameter), and in that sense scanning movements only provide an indirect measure of memory usage. We therefore chose to keep the basic model simple and not include any memory across multiple transitions, except for a rudimentary inhibition of return mechanism in the simulation of the eye movement version. The excellent fits generated by the model indeed indicate that a more complex, cumulative memory mechanism is not necessary. There may be good reasons for the cognitive system to not accumulate object memories across multiple transitions here. For example, when none of the remembered objects are found during a search round, one might as well forget them as they are less likely to be the target (whether that is truly the case or not). Forgetting objects may thus prevent unnecessarily long memory searches (cf. Wolfe, 2012). Note that the higher sampling cost in the mouse experiments was associated with both the constrained visual accessibility of objects (no attentional guidance) and the expensive motor cost of mouse movements (vs. eye movements). Future studies could explicitly separate the influence from the two factors by manipulating the amount of guidance available or by making eye movements "costlier." Another direction for future research would be to explore the factors that make it easier or harder to find a match. We speculate that the distance between pair members, but also their relative saliency, complexity, or meaningfulness may play a role. Finally, at a more general level, pair search is just one example of a fuzzy or abstract search goal. It would be worth studying other situations where the specific target is unknown, such as "find a present that will go well with Susan's new interior," or "who is not present at the meeting today."

In conclusion, when performing a search that requires dynamic exchanges of memory and perceptual input, observers adaptively balance the costs of external sampling and internal memorizing in accordance with task factors.

References

Alexander, R. G., & Zelinsky, G. J. (2011). Visual similarity effects in categorical search. *Journal of Vision*, *11*(8), Article 9. https://doi.org/10.1167/11.8.9

Alfandari, D., Belopolsky, A. V., & Olivers, C. N. (2019). Eye movements reveal learning and information-seeking in attentional template acquisition. *Visual Cognition*, *27*(5–8), 467–486. https://doi.org/10.1080/13506285.2019.1636918

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*(1), 66–80. https://doi.org/10.1162/jocn.1995.7.1.66

Becker, S. I., Folk, C. L., & Remington, R. W. (2010). The role of relational information in contingent capture. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(6), 1460–1476. https://doi.org/10.1037/a0020370

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14325–14329. https://doi.org/10.1073/pnas.0803390105

Bravo, M. J., & Farid, H. (2009). The specificity of the search template. *Journal of Vision*, *9*(1), Article 34. https://doi.org/10.1167/9.1.34

Carlisle, N. B., Arita, J. T., Pardo, D., & Woodman, G. F. (2011). Attentional templates in visual working memory. *Journal of Neuroscience*, *31*(25), 9315–9322. https://doi.org/10.1523/JNEUROSCI.1097-11.2011

Chan, L. K., & Hayward, W. G. (2013). Visual search. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(4), 415–429. https://doi.org/10.1002/wcs.1235

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Draschkow, D., Kallmayer, M., & Nobre, A. C. (2021). When natural behavior engages working memory. *Current Biology*, *31*(4), 869–874.e5. https://doi.org/10.1016/j.cub.2020.11.013

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433–458. https://doi.org/10.1037/0033-295X.96.3.433

Geng, J. J., DiQuattro, N. E., & Helm, J. (2017). Distractor probability changes the shape of the attentional template. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(12), 1993–2007. https://doi.org/10.1037/xhp0000430

Geng, J. J., & Witkowski, P. (2019). Template-to-distractor distinctiveness regulates visual search efficiency. *Current Opinion in Psychology*, *29*, 119–125. https://doi.org/10.1016/j.copsyc.2019.01.003

Hayhoe, M. M. (2017). Vision and action. *Annual Review of Vision Science*, *3*(1), 389–413. https://doi.org/10.1146/annurev-vision-102016-061437

Hayhoe, M. M., & Matthis, J. S. (2018). Control of gaze in natural environments: Effects of rewards and costs, uncertainty and memory in target selection. *Interface Focus*, *8*(4), Article 20180009. https://doi.org/10.1098/rsfs.2018.0009

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature*, *394*(6693), 575–577. https://doi.org/10.1038/29068

Horowitz, T. S., & Wolfe, J. M. (2001). Search for multiple targets: Remember the targets, forget the search. *Perception & Psychophysics*, *63*(2), 272–285. https://doi.org/10.3758/bf03194468

Hout, M. C., & Goldinger, S. D. (2010). Learning in repeated visual search. *Attention, Perception, & Psychophysics*, *72*(5), 1267–1282. https://doi.org/10.3758/APP.72.5.1267

Hout, M. C., & Goldinger, S. D. (2012). Incidental learning speeds visual search by lowering response thresholds, not by improving efficiency: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 90–112. https://doi.org/10.1037/a0023894

Hout, M. C., & Goldinger, S. D. (2015). Target templates: The precision of mental representations affects attentional guidance and decision-making in visual search. *Attention, Perception, & Psychophysics*, *77*(1), 128–149. https://doi.org/10.3758/s13414-014-0764-6

Hulleman, J., & Olivers, C. N. (2017). The impending demise of the item in visual search. *Behavioral and Brain Sciences*, *40*, Article e132. https://doi.org/10.1017/S0140525X15002794

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(3), 604–610. https://doi.org/10.1037/0096-1523.7.3.604

Kerzel, D. (2019). The precision of attentional selection is far worse than the precision of the underlying memory representation. *Cognition*, *186*, 20–31. https://doi.org/10.1016/j.cognition.2019.02.001

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*(4), 138–147. https://doi.org/10.1016/S1364-6613(00)01452-2

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558–78. https://doi.org/10.1037/a0019165

Li, C. L., Aivar, M. P., Tong, M. H., & Hayhoe, M. M. (2018). Memory shapes visual search strategies in large-scale environments. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-22731-w

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400. https://doi.org/10.1016/j.tics.2013.06.006

Maxfield, J. T., & Zelinsky, G. J. (2012). Searching through the hierarchy: How level of target categorization affects visual search. *Visual Cognition*, *20*(10), 1153–1163. https://doi.org/10.1080/13506285.2012.735718

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron*, *53*(4), 605–617. https://doi.org/10.1016/j.neuron.2007.01.018

Olivers, C. N., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences*, *15*(7), 327–334. https://doi.org/10.1016/j.tics.2011.05.004

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology*, *46*(3), 461–488. https://doi.org/10.1037/h0084327

Somai, R. S., Schut, M. J., & Van der Stigchel, S. (2020). Evidence for the world as an external memory: A trade-off between internal and external visual memory storage. *Cortex*, *122*, 108–114. https://doi.org/10.1016/j.cortex.2018.12.017

Standing, L. (1973). Learning 10000 pictures. *The Quarterly Journal of Experimental Psychology*, *25*(2), 207–222. https://doi.org/10.1080/14640747308400340

Witkowski, P., & Geng, J. J. (2019). Learned feature variance is encoded in the target template and drives visual search. *Visual Cognition*, *27*(5–8), 487–501. https://doi.org/10.1080/13506285.2019.1645779

Wolfe, J. M. (2012). Saved by a log: How do humans perform hybrid visual and memory search? *Psychological Science*, *23*(7), 698–703. https://doi.org/10.1177/0956797612443968

Wolfe, J. M. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, *28*(4), 1060–1092. https://doi.org/10.3758/s13423-020-01859-9

Wolfe, J. M., Klempen, N., & Dahlen, K. (2000). Postattentive vision. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 693–716. https://doi.org/10.1037//0096-1523.26.2.693

Woodman, G. F., & Chun, M. M. (2006). The role of working memory and long-term memory in visual search. *Visual Cognition*, *14*(4–8), 808–830. https://doi.org/10.1080/13506280500197397

Young, A. H., & Hulleman, J. (2013). Eye movements reveal how task difficulty moulds visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 168–190. https://doi.org/10.1037/a0028679

Yu, X., Hanks, T. D., & Geng, J. J. (2022). Attentional guidance and match decisions rely on different template information during visual search. *Psychological Science*, *33*(1), 105–120. https://doi.org/10.1177/09567976211032225

Yu, X., Johal, S. K., & Geng, J. J. (2022). Visual search guidance uses coarser template information than target-match decisions. *Attention, Perception, & Psychophysics*, *84*(5), 1–14. https://doi.org/10.3758/s13414-022-02478-3

Yu, X., Zhou, Z., Becker, S. I., Boettcher, S. E. P., & Geng, J. J. (in press). The good-enough theory of attentional guidance. *Trends in Cognitive Sciences*.

Received October 19, 2022
Revision received January 13, 2023
Accepted January 22, 2023 ■