Prospective search time estimates reveal the strengths and limits of internal models of visual search

Matan Mazor[1,2], Max Siegel[3], & Joshua B. Tenenbaum[3]

[1] Wellcome Centre for Human Neuroimaging, University College London

[2] Department of Psychological Sciences, Birkbeck, University of London

[3] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

In press, *Journal of Experimental Psychology: General*

Anonymized data, analysis scripts, experiment demos, and stimulus materials are available at github.com/matanmazor/metaVisualSearch. Experiments 2, 4, 5 and 6 were pre-registered prior to data collection. Pre-registration documents are available at https://osf.io/e48n5/registrations. M.M. was supported by a UCL Bogue Fellowship. This research was funded in whole, or in part, by the Wellcome Trust [Grant 203147/Z/16/]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The authors have no conflicting interests to disclose. We thank two anonymous reviewers for their useful comments and input, Noam and Bat-Sheva Mazor for help with developing the bonus structure, and Felix Henninger for assistance and guidance in programming the online experiments. We would especially like to thank signed reviewer Jeremy Wolfe for invaluable feedback on previous drafts of this paper, and for suggesting we run Experiments 5 and 6. Correspondence concerning this article should be addressed to Matan Mazor, WC1E 7HX, London UK. E-mail: mtnmzor@gmail.com.

Author note

Correspondence concerning this article should be addressed to Matan Mazor, WC1E 7HX, London UK. E-mail: mtnmzor@gmail.com

Abstract

Having an internal model of one's attention can be useful for effectively managing limited perceptual and cognitive resources. While previous work has hinted at the existence of an internal model of attention, it is still unknown how rich and flexible this model is, whether it corresponds to one's own attention or to a generic person-invariant schema, and whether it is specified as a list of facts and rules, or alternatively as a probabilistic simulation model. To this end, we tested participants' ability to estimate their own behavior in a visual search task with novel displays. In six online experiments (four pre-registered), prospective search time estimates reflected accurate metacognitive knowledge of key findings in the visual search literature, including the set-size effect, higher efficiency of color over conjunction search, and the asymmetric contributions of target and distractor identities to search difficulty. In contrast, estimates were biased to assume serial search, and demonstrated little to no insight into sizeable effects of search asymmetries for basic visual features, and of visual and semantic similarity between target and distractors. Together, our findings reveal a complex picture, where internal models of visual search are sensitive to some, but not all, of the factors that make some searches more difficult than others.

Keywords: metacognition, self-model, attention-schema, visual search

Word count: 9686

Prospective search time estimates reveal the strengths and limits of internal models of visual search

Introduction

In order to efficiently interact with the world, agents construct *mental models*: simplified representations of the environment and of other agents that are accurate enough to generate useful predictions and handle missing data (Forrester, 1971; Friston, 2010; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). For example, participants' ability to predict the temporal unfolding of physical scenes has been attributed to an 'intuitive physics engine': a simplified model of the physical world that uses approximate, probabilistic simulations to make rapid inferences (Battaglia, Hamrick, & Tenenbaum, 2013). Similarly, having a simplified model of planning and decision-making allows humans to infer the beliefs and desires of other agents based on their observed behavior (Baker, Saxe, & Tenenbaum, 2011). Finally, in motor control, an internal model of one's motor system and body allows subjects to monitor and control their body (Wolpert, Ghahramani, & Jordan, 1995). This internal forward model has also been proposed to play a role in differentiating self and other (Blakemore, Wolpert, & Frith, 1998). In recent years, careful experimental and computational work has advanced our understanding of these internal models: their scope, the abstractions that they make, and the consequences of these abstractions for faithfully and efficiently modeling the environment.

Agents may benefit from having a simplified model not only of the environment, other agents, and their motor system, but also of their own perceptual, cognitive and psychological states. For example, it has been suggested that knowing which items are more subjectively memorable is useful for making negative recognition judgments ("I would have remembered this object if I saw it," Brown, Lewis, & Monk, 1977). Similarly, children guided their decisions and evidence accumulation based on model-based expectations about the perception of hidden items (Siegel, Magid, Pelz, Tenenbaum, & Schulz, 2021). In the context of perception and attention, Graziano and Webb (2015) argued that having a simplified Attention Schema — a simplified model of attention and its dynamics — is crucial for monitoring and controlling one's attention, similar to how a body-schema supports motor control.

Indeed, people are not only capable of predicting the temporal unfolding of physical scenes, or the behavior of other agents, but also the workings of their own attention under hypothetical scenarios. In one study, participants held accurate beliefs about the serial nature of visual search for a conjunction of features, and the parallel nature of visual search for a distinct color (Levin & Angelone, 2008). Similarly, the majority of third graders knew that the addition of distractors makes finding the target harder, particularly if the distractors and target are of the same color (Miller & Bigi, 1977). These and similar studies established the existence of metacognitive knowledge about visual search, as a result raising new questions about its structure, limits, and origins. We identify three such open questions. First, do internal models of visual search represent search difficulty along a gradient, or alternatively classify search displays as being either parallel or serial? Second, to what extent is knowledge about visual search learned or calibrated based on first-person experience? And third, are internal models of visual search structured as a list of facts and laws, or as an approximate probabilistic simulation?

Here we take a first step toward providing answers to these three questions, using visual search as our model test case for internal models of perception and attention more generally. Participants estimated their prospective search times in visual search tasks and then performed the same searches. Similar to using colliding balls (Smith & Vul, 2013) and falling blocks (Battaglia et al., 2013) to study intuitive physics, here we chose visual search for being thoroughly studied and for following robust behavioral laws. In Experiments 1 and 2, we used simple colored shapes as our stimuli, and compared participants' internal models to scientific theories of attention that distinguish parallel from serial processing. We found that participants represented the relative efficiency of different search tasks, but had a persistent bias to assume serial search. In Experiments 3 and 4 we used unfamiliar stimuli from the Omniglot dataset (Lake, Salakhutdinov, Gross, & Tenenbaum, 2011) with the purpose of testing the richness and compositional nature of participants' internal models, and their reliance on person-specific knowledge. We find that participants are capable of predicting their search times, even for novel stimuli. Furthermore, we show that for complex stimuli, internal models of visual search are better fitted to one's own search behavior compared with the search behavior of other participants. Finally, in Experiments 5 and 6, we find important limitations of these models: they fail to represent asymmetries between searching for the presence or absence of basic visual features, and are blind to the effects of semantic target-distractor similarity on search difficulty. Together, we find that people are capable of estimating the relative search difficulty of previously unseen searches, but that this ability is limited by having only partial insight into the many factors that affect visual search.

Experiments 1 and 2: shape, orientation, and color

An internal model of visual search may take a similar form to that of a scientific theory, by specifying an ontology of concepts and a set of causal laws that operate over them (Gerstenberg & Tenenbaum, 2017; Gopnik & Meltzoff, 1997). For example, participants may hold an internal model of visual search that is similar to Anne Treisman's *Feature Integration Theory*. According to this theory, visual search comprises two stages: a pre-attentive parallel stage, and a serial focused attention stage (Treisman, 1986; Treisman & Sato, 1990). In the first stage, visual features (such as color, orientation, and intensity) are extracted from the display to generate spatial ‘feature maps’. Targets that are defined by a single feature with respect to their surroundings can be located based on these feature maps alone (*feature search*; for example searching for a red car in a road full of yellow taxis). Since the extraction of a feature map is pre-attentive, in these cases search can be completed immediately. In contrast, sometimes the target can only be identified by integrating over multiple features (*conjunction search*; for example if the road has not only yellow taxis, but also red buses). In such cases, attention must be serially deployed to items in the display until the target is identified.

A simplifying assumption of Feature Integration Theory is that there is no transfer of information between the pre-attentive and focused attention stages. In other words, observers cannot selectively direct their focused attention to items that produced strong activations in the pre-attentive stage. *Guided Search* models (Wolfe, 1994, 2021; Wolfe, Cave, & Franzel, 1989) assume instead that participants use these pre-attentive guiding signals in their serial search. Compared to Feature Integration Theory, Guided Search models provide much better fit to empirical data, at the expense of being more complex and rich in detail. To date, it is unknown where internal models of visual search fall on this performance-complexity trade-off: do people differentiate between parallel and serial searches like in Feature Integration Theory, or do they represent search difficulty on a continuum, more like Guided Search?

In Experiments 1 and 2 we used stimuli that lend themselves to a categorical distinction between parallel and serial search: simple geometrical shapes of different colors and orientations. We asked whether participants’ internal models of visual search predict which search displays demand serial deployment of attention and which don’t. Critically, participants gave their search time estimates before they were asked to perform searches involving these or similar stimuli, so their search time estimates reflected prior beliefs about search efficiency. Experiment 2 was designed to replicate and generalize the results of Exp. 1 to a new stimulus dimension (orientation) and distractor set sizes. Our hypotheses and analysis plan for Experiment 2, based on the results of Experiment 1, were pre-registered prior to data collection (pre-registration document: doi.org/10.17605/OSF.IO/2DPQ9). Raw data, experiment demos, and full analysis scripts are available at github.com/matanmazor/metaVisualSearch.

Participants

Experiments were approved by the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects under protocol 0812003014. All participants gave their informed consent prior to participating. For Exp. 1, 100 participants were recruited from Amazon's crowdsourcing web-service Mechanical Turk. Exp. 1 took about 20 minutes to complete. Each participant was paid $2.50. The highest performing 30% of participants received an additional bonus of $1.50. For Exp. 2, 100 participants were recruited from the Prolific crowdsourcing web-service. The experiment took about 15 minutes to complete. Each participant was paid £1.5. The highest performing 30% of participants received an additional bonus of £1.

Procedure

The study was built using the Lab.js platform (Henninger, Shevchenko, Mertens, Kieslich, & Hilbig, 2019) and hosted on a JATOS server (Lange, Kühn, & Filevich, 2015). Demo versions of all four experiments are available at github.com/matanmazor/metaVisualSearch.

Familiarization. First, participants were acquainted with the visual search task. The instructions for this part were as follows:

In the first part, you will find a target hidden among distractors. First, a gray cross will appear on the screen. Look at the cross. Then, the target and distractors will appear. When you spot the target, press the spacebar as quickly as possible. Upon pressing the spacebar, the target and distractors will be replaced by up to 5 numbers. To move to the next trial, type in the number that replaced the target.

The instructions were followed by four trials of an example visual search task (searching for a *T* among 7 *L*s). Feedback was delivered on speed and accuracy. The purpose of this part of the experiment was to familiarize participants with the task.

Estimation. After familiarization, participants estimated how long it would take them to perform various visual search tasks involving novel stimuli and various set sizes. On each trial, they were presented with a target stimulus and a display of distractors and were asked to estimate how long it would take to find the target if it was hidden among the distractors (see Fig. 1).

To motivate accurate estimates, we explained that these visual search tasks will be performed in the last part of the experiment, and that bonus points will be awarded for trials in which participants detect the target as fast or faster than their search time estimate. The number of points awarded for a successful search changed as a function of the estimate given for the same search, such that more points were offered for riskier estimates. In order to meaningfully compare estimates for different searches, it was important that any tendency to produce risky or conservative estimates is conserved across all searches. To achieve that, the number of points offered for a successful search was set to 10/√estimate. We chose this rule because for right-skewed log-normal reaction time distributions, an optimal strategy is to consistently choose an estimate that is aligned with the 70th quantile of the estimated RT distribution (see Appendix). The report scale ranged from 0.1 to 4 seconds.

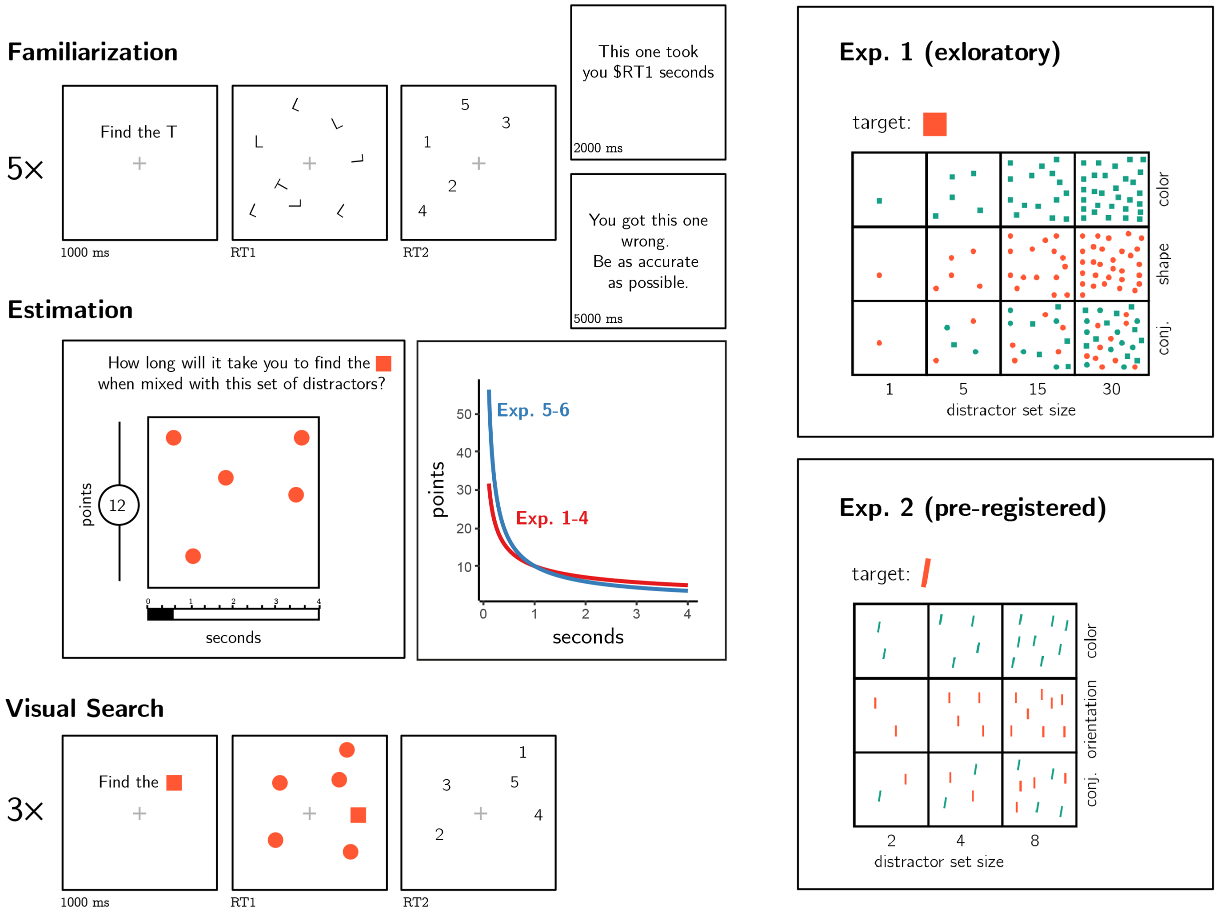


Figure 1. Experimental design. Participants first performed five similar visual search trials and received feedback about their speed and accuracy. Then, they were asked to estimate the duration of novel visual search tasks. Bonus points were awarded for accurate estimates, and more points were awarded for risky estimates. Finally, in the visual search part participants performed three consecutive trials of each visual search task for which they gave a search time estimates. Right panels: stimuli used for Experiments 1 and 2.

After one practice trial (estimating search time for finding one *T* among 3 randomly positioned *L*s), we turned to our stimuli of interest. In Experiment 1, participants estimated how long it would take them to find a red (#FF5733) square among green (#16A085) squares (color condition), red circles (shape condition) and a mix of green squares, red circles, and green circles (shape-color conjunction condition), for set sizes 1, 5, 15 and 30. Together, participants estimated the expected search time of 12 different search tasks (see Figure 1, upper right panel). In

Experiment 2, participants rated how long it would take them to find a red tilted bar (20° off vertical) among green titled bars (color condition), red vertical bars (orientation condition) and a mix of green tilted and red vertical bars (orientation-color conjunction condition) for set sizes 2, 4, and 8. Together, participants estimated the expected search time of 9 different search tasks (see Figure 1, lower right panel). In both experiments, the order of estimation trials was randomized between participants.

Visual Search. Participants performed three consecutive search tasks for each of the 12 (Exp. 1) or 9 (Exp. 2) search types. The order of presentation was randomized between participants. No feedback was delivered about speed. To motivate accurate responses, error trials were followed by a 5-second pause.

Results

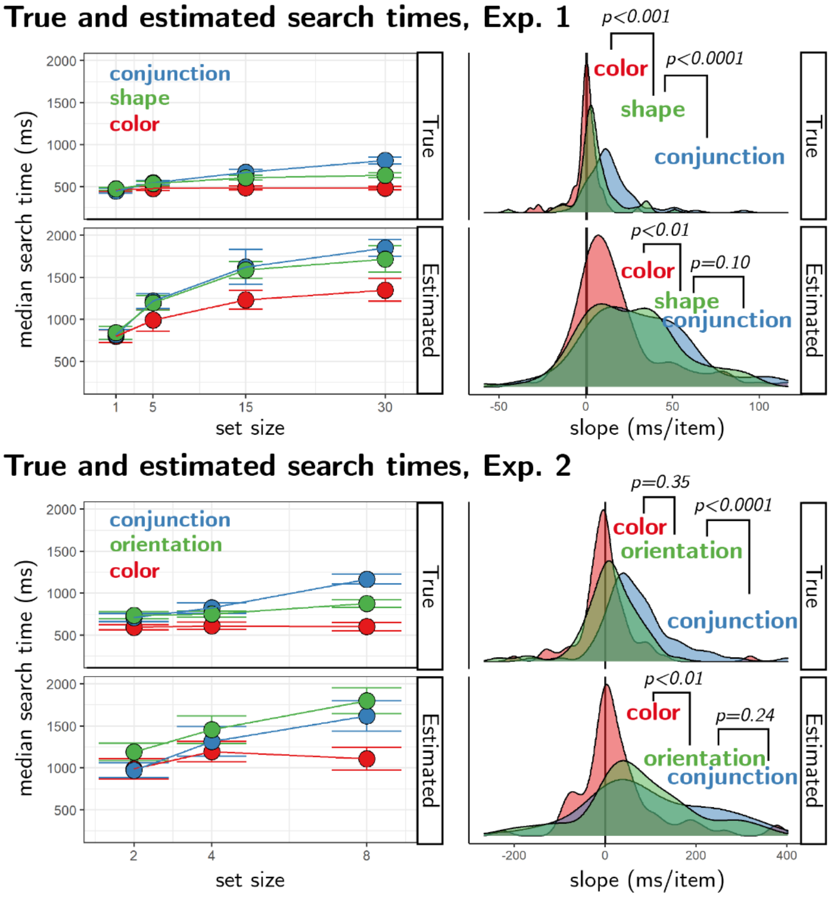
Accuracy in the visual search task was reasonably high in both Experiments (Exp. 1: $M = 0.93$, 95% CI $[0.90,0.96]$; Exp. 2: $M = 0.82$, 95% CI $[0.77,0.87]$). Error trials and visual search trials that took shorter than 200 milliseconds or longer than 5 seconds were excluded from all further analysis. Participants were excluded if more than 30% of their trials were excluded based on the aforementioned criteria, leaving 89 and 74 participants for the main analysis of Experiments 1 and 2, respectively.

Search times. For each participant and distractor type, we extracted the slope of the function relating RT to distractor set size. As expected, search slopes for color search were not significantly different from zero in Exp. 1 (-0.40 ms/item; $t(88) = -0.45, p = .652$, $\text{BF}_{01} = 7.74$) and Exp. 2 (0.51 ms/item; $t(73) = 0.07, p = .946$, $\text{BF}_{01} = 7.80$). This is consistent with color being a basic feature that is not dependent on serial attention for its extraction by the visual system (Treisman, 1986; Treisman & Sato, 1990). The slope for shape search was close, but significantly higher than zero (5.66 ms/item; $t(88) = 4.35, p < .001$), and the slope for orientation was numerically higher than zero (11.05 ms/item) but not significantly so ($t(73) = 1.50, p = .139$, $\mathrm{BF}_{01} = 2.70$). In both Experiments, conjunction search gave rise to search slopes significantly higher than zero (Exp. 1: 14.80 ms/item ($t(88) = 9.16, p < .001$; Exp. 2: 72.14 ms/item ($t(73) = 7.50, p < .001$; see Figure 2).

Estimation accuracy. We next turned to analyze participants' prospective search time estimates, and their alignment with actual search times. In both Experiments, participants generally overestimated their search times. This was the case for all search types across the two Experiments (see Figure 2, left panels). This is expected, based on our bonus scheme that incentivized conservative estimates (see Appendix). Despite this bias, estimates were correlated with true search times, supporting a metacognitive insight into visual search behavior (see Fig. 2, left panels. Within subject Spearman correlations, Exp. 1: $M = 0.28$, 95% CI $[0.21, 0.35]$, $t(88) = 7.77, p < .001$; Exp 2: $M = 0.16$, 95% CI $[0.07, 0.26]$, $t(73) = 3.48, p = .001$).

To test participants' internal models of visual search, we analyzed their estimates as if they were search times, and extracted *estimation slopes* relating estimates to the number of distractors in the display (see Fig. 2, right panels). Estimation slopes (expected ms/item) were steeper than search slopes for all search types. In particular, although search time for a deviant color was unaffected by the number of distractors, participants estimated that color searches with more distractors should take longer (mean estimated slope in Exp. 1: 17.76 ms/item; $t(88) = 6.35, p < .001$; in Exp 2: 29.43 ms/item; $t(73) = 2.63, p = .010$). In other words, at the group level, participants showed no metacognitive insight into the parallel nature of color search.

Although they were significantly different from zero, in both Experiments estimation slopes for color search were significantly shallower than for conjunction search (Exp. 1: $t(88) = 4.08$, $p < .001$, Exp. 2: $t(73) = 3.87$, $p < .001$). In contrast, although true search slopes were shallower for shape and orientation than for conjunction (p’s<0.001), the difference in estimation slopes was not significant (difference between shape and conjunction slopes: $t(88) = 1.65$, $p = .103$; difference between orientation and conjunction slopes: $t(73) = 1.18$, $p = .244$).


True and estimated search times, Exp. 1

True and estimated search times, Exp. 2

Figure 2. Results from Experiments 1 and 2. Left: median actual and estimated search times as a function of set size for the different search types (coded by color). Error bars represent the standard error of the median, estimated with bootstrapping. Right panels: distribution of search slopes for actual and estimated search types.

Experiments 3 and 4: complex, unfamiliar stimuli

In Experiments 1 and 2 an internal model of visual search allowed participants to accurately estimate how long it would take them to find a target stimulus in arrays of distractor stimuli. Participants had insight into the set-size effect and into the fact that conjunction searches are more difficult than color searches. Positive color search slopes that are nevertheless significantly shallower than conjunction search slopes further suggested a graded representation of search efficiency, but no awareness of the possibility of parallel processing of preattentive basic features. An alternative interpretation is that a gradient of positive search slopes emerges due to a group averaging effect of individual dichotomous representations. If some participants represent color search as parallel, and others as equally difficult as conjunction search, the mean slope for color search would be higher than zero and significantly lower than for conjunction search.

In Experiments 3 and 4 we addressed this possibility, and further asked how rich this model is, by using displays of complex stimuli with which participants are unlikely to have any prior experience (letters from a medieval Alphabet and from the Futurama TV series, hand drawn by Mechanical Turk workers). Here, insight into the set size effect and its absence in feature searches would not be useful for generating accurate search time estimates. Instead, participants' internal model of visual search must be capable of extracting relevant features from rich stimuli, and using these features to generate graded stimulus-specific predictions. Using these more complex stimuli further allowed us to ask if search-time estimates rely on person-specific knowledge, as subjects are expected to vary more in their search behavior in more complex displays. Exp. 4 followed Exp. 3 and was pre-registered (pre-registration document:

doi.org/10.17605/OSF.IO/DPRTK). Raw data,experiment demos, and full analysis scripts are available at github.com/matanmazor/metaVisualSearch.

Participants

For Exp. 3, 100 participants were recruited from the Prolific crowdsourcing web-service. The experiment took about 15 minutes to complete. Participants were paid £1.5. The highest performing 30% of participants received an additional bonus of £1. For Exp. 4, 200 participants were recruited from the Prolific crowdsourcing web-service. We recruited more participants for Exp. 4 in order to have sufficient statistical power for our inter-subject correlation analysis. The experiment took about 8 minutes to complete. Participants were paid $1.27. The highest performing 30% of participants received an additional bonus of $0.75.

Procedure

The procedure for Experiments 3 and 4 was similar to that of Exp. 1 with several changes.

Stimuli were letters drawn by Mechanical Turk workers (Lake et al., 2011), instead of geometrical shapes (see Fig. 3). In Exp. 3, we used letters from the *Alphabet of the Magi*. In Exp. 4, we used letters from the *Futurama* television series as well as Latin letters. We explained to participants that they will search for a specific letter (the target letter) among copies of another letter (the distractor letter). In Exp. 3, both target and distractor were letters from the Alphabet of the Magi, and distractors were drawn by different Mechanical Turk workers. In Exp. 4, on half of the trials the target was a Latin letter and distractors were Futurama letters and on the other half the target was a Futurama letter and distractors were Latin letters. In these experiments, distractors were copies of the same letter drawn by the same Mechanical Turk worker. This was important for our visual search asymmetry analysis (see below).

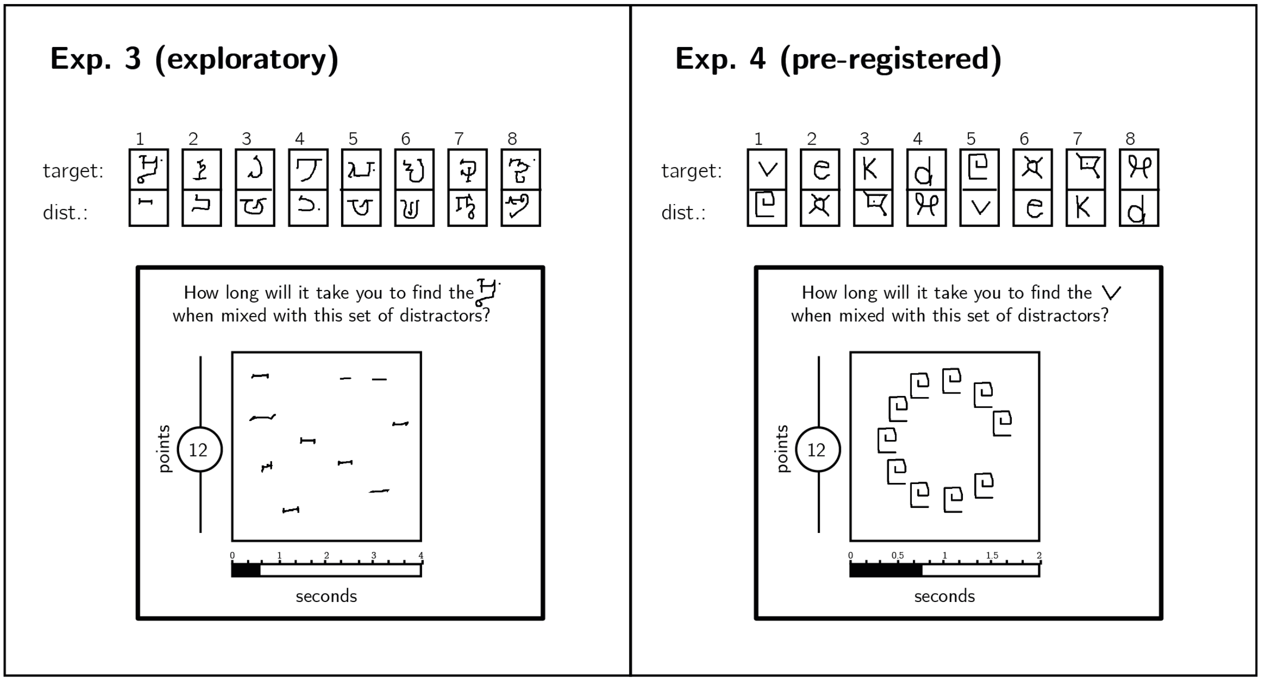

Exp. 3 (exploratory)
1 2 3 4 5 6 7 8
target:
dist.:
How long will it take you to find the when mixed with this set of distractors?
points
12
0 1 2 3 4
seconds
Exp. 4 (pre-registered)
1 2 3 4 5 6 7 8
target:
dist.:
How long will it take you to find the V when mixed with this set of distractors?
points
12
0 0.5 1 1.5 2
seconds

Figure 3. In Exp. 3, stimuli were characters from the Alphabet of the Magi, and distractors were drawn by different Mechanical Turk users. In Exp.4, stimuli were characters from the Latin and Futurama alphabets. Stimulus pairs 1-4 and 5-8 are identical except for the target assignment. In Exp. 4, all distractors in a display were drawn by the same Mechanical Turk user, and were presented on an invisible clock face.

In the familiarization part, we used as target and distractors two letters from the Alphabet of the Magi (Exp. 3) and two letters from the Futurama alphabet (Exp. 4). Importantly, these letters were only used for training, and did not appear in the Estimation or Visual search parts. In the Estimation part participants gave search time estimates for 8 search tasks, all involving 10 distractors, and in the Visual Search part they performed these search tasks. To minimize random variation in spatial configurations (which was important for the search asymmetry analysis), in Exp. 4 letters appeared on an invisible clock face. Finally, the report scale ranged from 0.1 to 4 seconds in Exp. 3 and to 2 seconds in Exp. 4.

Results

Accuracy in the visual search task was high in both experiments (Exp. 3: $M = 0.89$, 95% CI $[0.86,0.92]$; Exp. 4: $M = 0.97$, 95% CI $[0.96,0.98]$). Error trials and visual search trials that took shorter than 200 milliseconds or longer than 5 seconds were excluded from all further analysis. Participants were excluded if more than 30% of their trials were excluded based on the aforementioned criteria, leaving 88 and 200 participants for the main analysis of Experiments 3 and 4, respectively.

Estimation accuracy. In both experiments, search time estimates were positively correlated with true search times (within-subject Spearman correlations in Exp. 3: $M = 0.44$, 95% CI $[0.37,0.52]$, $t(86) = 12.16, p < .001$; Exp. 4: $M = 0.10$, 95% CI $[0.05,0.15]$, $t(191) = 3.67, p < .001$; see Figures 4 and 6A). The correlation between search time and search time estimates was significantly weaker in Experiment 4 ($\Delta M = 0.35$, 95% CI $[0.26,0.43]$, $t(181.02) = 7.60, p < .001$)). This difference in correlation strength is likely the result of a narrower range of search times in Exp. 4 (with median search times 566 - 684 ms, per display) than in Exp. 3 (649 - 1615 ms), increasing the relative contribution of measurement noise to search times, and attenuating correlations as a result. Indeed, the mean Spearman correlation between the search times of a given participants and the median search times of all other participants, a measure of the noisiness of the data that is independent of search time estimates, dropped from 0.74 in Exp. 3 to 0.29 in Exp. 4 ($t(284.57) = 14.36, p < .001$).

Importantly, in both experiments all searches involved exactly 10 distractors, so a positive correlation could not be driven by the effect of distractor set size. Furthermore, since participants had no prior experience with our stimuli, their estimates could not have been informed by explicit knowledge about specific letters (‘The third letter in the *Alphabet of the Magi* pops out to attention when presented between instances of the fourth letter', or 'the fifth letter in the *Futurama Alphabet* is difficult to find when presented among *d*s'). These positive correlations reveal a more intricate knowledge of visual search. Our next two analyses were designed to test whether estimates were based on person-specific knowledge, and whether their generation involved a simulation of the search process.

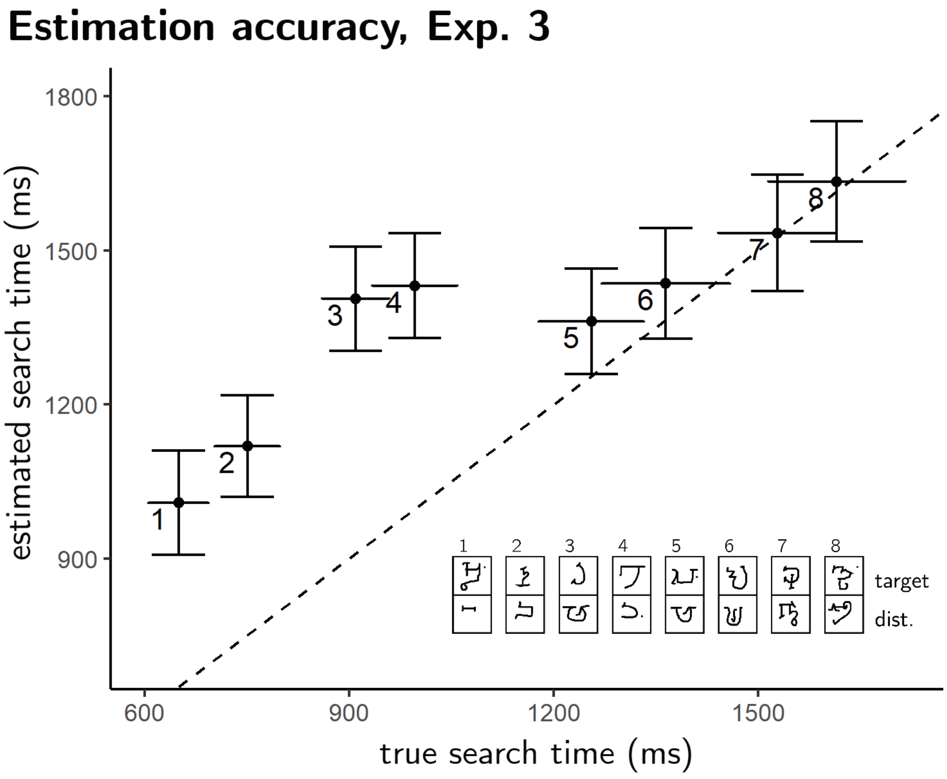


Figure 4. Estimated search times plotted against true search times in Experiment 2. The dashed line indicates $y = x$. Legend: each search task involved searching for one Omniglot character (top letter) among ten tokens of a second Omniglot character, drawn by 10 different MTurk workers (bottom letter).

Cross-participant correlations. We chose unfamiliar letters as stimuli for Experiments 3 and 4 in order to make heuristic-based estimation more difficult, and to encourage an introspective estimation process. If participants were using idiosyncratic knowledge about their own attention, we would expect to find higher correlations between their search time estimates and their own search times (*self-self alignment*), compared to with the search times of a random surrogate participant (*self-other alignment*). To test this, we ran a non-parametric permutation test, comparing self-self and self-other alignment in prospective search time estimates. In Exp. 3, a numerical difference between self-self (mean Spearman correlation $M_r = 0.44$) and self-other alignment ($M_r = 0.41$) was marginally significant ($p_{perm} = 0.05$). In Experiment 4, we pre-registered this analysis and found a significant advantage for self-self alignment compared with self-other alignment (see Fig. 5; mean Spearman correlations for self-self $M_r = 0.10$ and self-other $M_r = 0.04$, $p_{perm} = 0.01$). This result can be interpreted as indicating that at least some of participants' internal model of visual search builds on idiosyncratic knowledge about their own attention. Alternatively, it may reflect inter-individual differences in the perception of complexity and similarity of targets and distractors. We unpack some implications of these two competing accounts in the General Discussion.

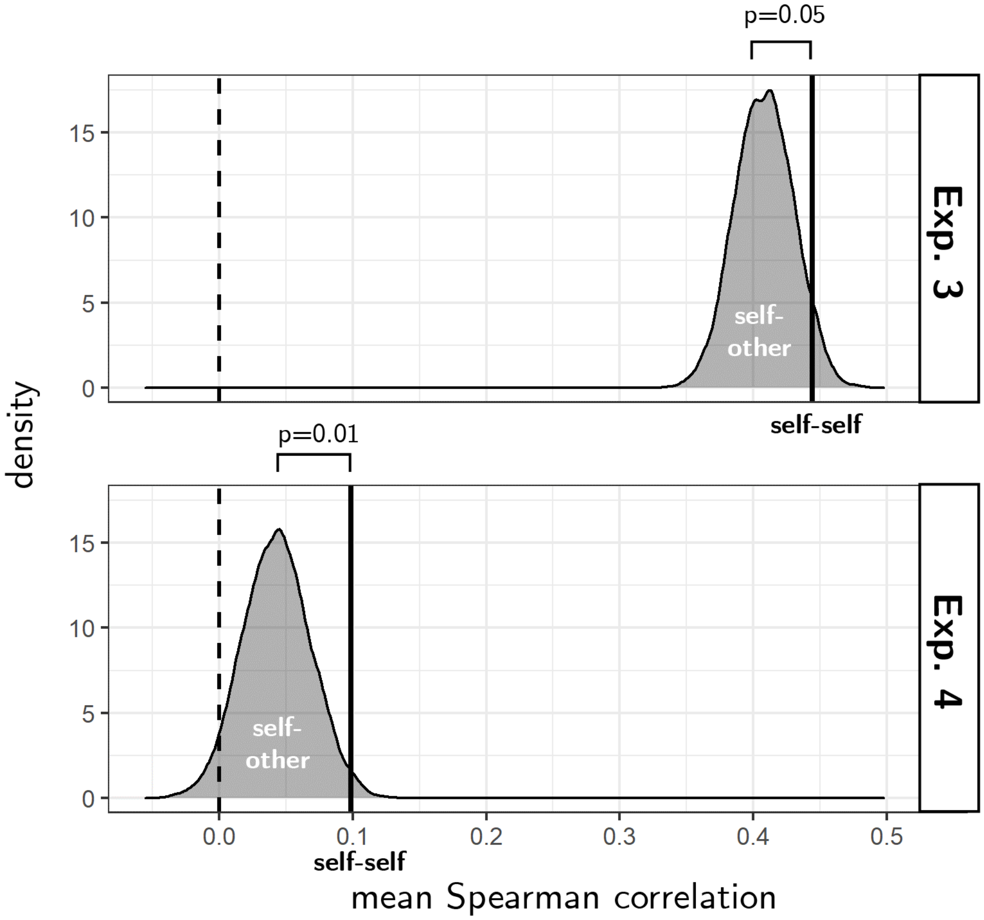


Figure 5. True correlation between estimates and search times (self-self alignment, vertical lines) plotted against a null distribution of correlations, when matching the estimates of each participant with the search time of a random surrogate participant (self-other alignment).

Estimation time. We next looked at the time taken to produce search time estimates in the Estimation part. We reasoned that if participants had to mentally simulate searching for the target in order to generate their search time estimates, they would take longer to estimate that a search task will terminate after 1500 compared to 1000 milliseconds. This is similar to how a linear alignment between the degree of rotation and response time in a mental rotation task was taken as support for an internal simulation that evolves over time (Shepard & Metzler, 1971). We find no evidence for within-subject correlation between estimates and the time taken to deliver them, not in Exp. 3 ($t(86) = 0.40, p = .692$) and not in Exp. 4 ($t(191) = 0.74, p = .458$). However, given that estimation times were three times longer than search time estimates (median time to estimate = 5 seconds in Exp. 3 and 3 seconds in Exp. 4), a simulation-driven correlation may have been masked by other factors that contributed to estimation times, such as motor control over the report slider.

Visual search asymmetry. To keep things simple, internal models of visual search may make the simplifying assumption that target and distractor stimuli contribute to search difficulty in similar ways. For example, models can specify that search time generally inversely scales with the perceived similarity between the target and distractor stimuli, without taking into account the different roles target and distractor play in determining search difficulty. Alternatively, internal models of visual search may represent the asymmetric nature of visual search tasks (finding an A among Bs is not the same as finding a B among As) at the expense of additional model complexity.

To test whether internal models of visual search were sensitive to the assignment of stimuli to target and distractor roles, we leveraged a well-established phenomenon in visual search: subjects are generally faster detecting an unfamiliar stimulus in an array of familiar distractors compared to when the target is familiar and the distractors are not (Malinowski & Hübner, 2001; Shen & Reingold, 2001; Zhang & Onyper, 2020). This asymmetry cannot be captured by a model of visual search that is blind to the assignment of stimuli to target and distractor roles. In Exp. 4, participants were presented with pairs of familiar and unfamiliar letters, and estimated their search time for finding the familiar letter among unfamiliar distractors and vice versa. This allowed us to test for visual search asymmetries in search times and in search time estimates.

As expected, searching for a familiar target among unfamiliar distractors was more difficult on average, with a difference of 41 milliseconds in search time ($t(199) = 4.41, p < .001$). To test if subjects were sensitive to the assignment of stimuli to target and distractor roles, we extracted individual subjects' Spearman correlations between search times and their reciprocal estimates (that is, the estimate for the same search with the target and distractor roles inverted). For example, instead of comparing search times for finding the letter v among 10 square spiral letters (stimulus pair 1) with estimates for the same search, we compared it with estimates for finding one square spiral letter among 10 v's (stimulus pair 5). If estimates were affected by the assignment of stimuli to target and distractor, this inversion should attenuate the correlation, but if visual search estimates reflected a symmetric notion of similarity the correlation should not be affected.

Inverting the target/distractor assignment dropped the correlation between estimates and search time to zero ($M = -0.01$, 95% CI $[-0.06,0.04]$), significantly lower than the original correlation ($M_D = 0.10$, 95% CI $[0.03,0.18]$, $t(191) = 2.63$, $p = .009$; see Fig. 6B). This is in contrast to what is expected if search time estimates reflected symmetric similarity judgments, and in line with an interpretation of our findings as evidence for an internal model of visual search that is sensitive to the assignment of stimuli to target or distractor roles.

Interestingly, however, a difference in mean estimated search time between familiar and unfamiliar targets did not reach statistical significance ($M = 9.88$, 95% CI $[-5.90,25.66]$, $t(199) = 1.23$, $p = .218$). A drop in subject-specific Spearman correlations without a significant difference in mean search times indicates that subjects' sensitivity to the assignment of stimuli to target and distractor roles was not fully captured by the metacognitive insight that familiar targets are more difficult to find. Subjects may have been sensitive to other visual properties that contributed to search asymmetries. In Exp. 5, we further explore sensitivity to three such features that produce robust asymmetries in visual search behavior: orientation, open edges, and addition of line strokes.

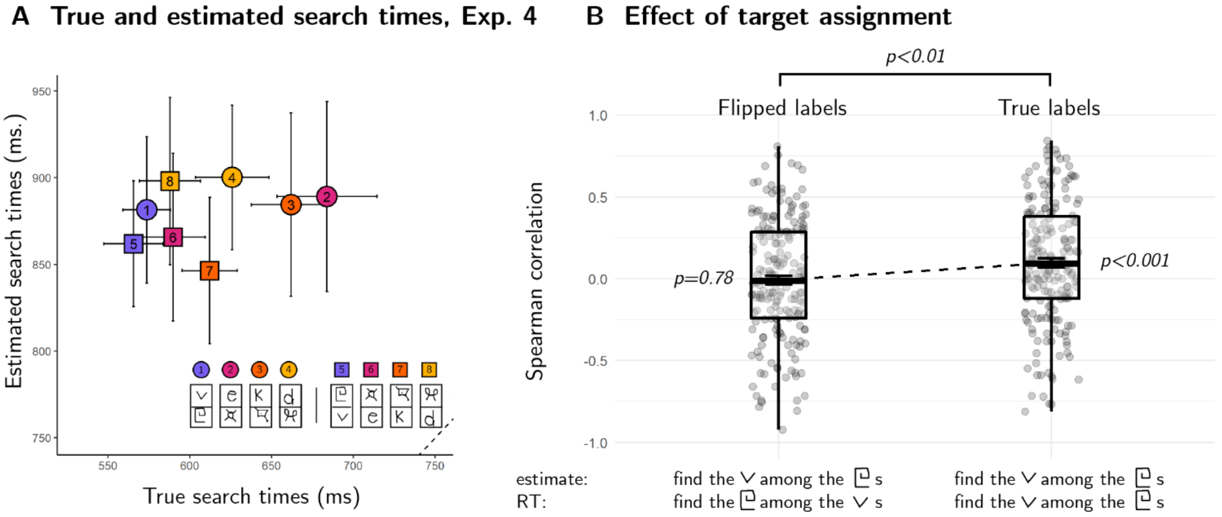


Figure 6. A. Median estimated search times plotted against true search times in Exp. 4. The dashed line indicates y=x. Legend: each search task involved searching for one character (top letter) among ten tokens of a different character (bottom letter). In four searches, the target character was from the Latin alphabet (circles), and in the other four from the Futurama alphabet (squares). Search pairs that involved the same pair of stimuli with opposite roles are marked by the same color. B. Spearman correlations between estimates and search times for true and target-distractor flipped labels in Exp. 4. Spearman correlations significantly dropped, indicating that participants were aware of the effect of target assignment on search time.

Experiment 5: three search asymmetries

In Exp. 4, search time estimates were sensitive to the assignment of stimuli to target and distractor roles, but not to the visual search asymmetry for familiar and unfamiliar stimuli. In Exp. 5, we examined three additional search asymmetries (line orientation, open edges, and line addition), and asked whether they are accurately specified in participants' internal models of visual search. Exp. 5 was pre-registered (pre-registration document: doi.org/10.17605/OSF.IO/VJQ2F). Raw data, experiment demos, and full analysis scripts are available at github.com/matanmazor/metaVisualSearch.

Participants

For Exp. 5, 203 participants were recruited from the Prolific crowdsourcing web-service. The experiment took about 10 minutes to complete. Participants were paid $1.59. The highest performing 30% of participants received an additional bonus of $1.59.

Procedure

The procedure for Experiments 5 was similar to that of Exp. 1 with several changes.

Participants estimated their prospective search times for three stimulus pairs. Within each pair, participants provided estimates for two versions of the search: one where the first stimulus serves as a target and the second as the distractor, and one where the roles were reversed. For each search, subjects provided estimates for set sizes of 6 and 18. The three stimulus pairs were 1) a vertical line and a tilted (20° off vertical) line, 2) a circle and a circle intersected by a line, and 3) a circle and a circle with an open gap (see Fig. 7, left panel). For brevity, we refer to these last stimuli as O, Q, and C. All three stimulus pairs have been shown to produce asymmetries in visual search time, such that the assignment of stimuli to target or distractor roles affects search time (Treisman & Gormican, 1988; Treisman & Souther, 1985).

The estimation scale ranged from 0 to 2 seconds. In Exp. 5, we adapted the estimate-to-points conversion rule to be $10/estimate^{3/4}$ rather than $10/estimate^{1/2}$. Making the number of offered points decline faster ensured that the optimal strategy is to report the median of the posterior distribution over reaction times, making it possible to directly compare median search times and prospective estimates.

In the visual search part, participants performed five consecutive instances of each search. In order to prevent subjects from relying on their iconic memory to identify the position of the target after making an initial response, stimuli were masked by a random black and white image for a duration of 50 milliseconds following spacebar responses.

Results

Accuracy in the visual search task was high ($M = 0.96$, 95% CI $[0.95, 0.97]$). Error trials and visual search trials that took shorter than 200 milliseconds or longer than 5 seconds were excluded from all further analysis. Participants were excluded if more than 30% of their trials were excluded based on the aforementioned criteria, leaving 200 participants for the main analysis of Experiment 5.

Visual search asymmetries. Search slopes were significantly different for the six searches ($F(3.57, 704.01) = 152.88$, $MSE = 1{,}167.92$, $p < .001$, $\hat{\eta}^2_G = .376$), with orientation search slopes shallower on average than the other two searches ($M = -139.41$, 95% CI $[-149.80, -129.01]$, $t(197) = -26.44$, $p < .001$).

Within the three stimulus pairs, we observed the expected search asymmetries. The mean search slope for finding one vertical target among multiple tilted distractors (18.12 ms/item) was significantly steeper than the slope for the inverse search (5.16 ms/item; $M = 12.97$, 95% CI $[8.27, 17.68]$, $t(197) = 5.43$, $p < .001$). Similarly, the mean search slope for finding one O target among multiple distractor Qs (61.78 ms/item) was significantly steeper than the slope for the inverse search (8.33 ms/item; $M = 53.44$, 95% CI $[47.66, 59.23]$, $t(199) = 18.22$, $p <$

.001). Finally, the mean search slope for finding one O target among multiple distractor Cs (59.88 ms/item) was significantly steeper than the slope for the inverse search (20.55 ms/item; $M = 39.80$, 95% CI $[33.30,46.31]$, $t(198) = 12.06, p < .001$).

Estimation accuracy. The mean Spearman correlation between search slopes and their estimates was 0.32 and significantly different from zero ($M = 0.32$, 95% CI $[0.28,0.36]$, $t(197) = 15.76, p < .001$). Contrary to our findings from Exp. 1-4, the average search estimate slope (17.48 ms/item) was significantly *shallower* than the average search slopes (29.00 ms/item; $M_D = -11.39$, 95% CI $[-14.76,-8.02]$, $t(197) = -6.66, p < .001$). This difference may be driven by the change to our estimate-to-points conversion rule, which now incentivized more risky estimates.

Overall, participants integrated information about the assignment of stimuli to target or distractor roles in providing their estimates. The mean Spearman correlation between search times and the estimates of their reciprocal searches (that is, searches with the same stimuli and set sizes, but an opposite target/distractor assignment) was 0.22 – significantly lower than the correlation between search times and their corresponding (non-reciprocal) estimates: 0.32 ($M = 0.10$, 95% CI $[0.05,0.15]$, $t(197) = 4.14, p < .001$).

However, when examining the effect on search slope within specific stimulus pairs, we found little to no support for asymmetries in prospective search time estimates. Estimation slopes were not sensitive to the search asymmetry for line orientation ($M = -0.68$, 95% CI $[-4.31,2.96]$, $t(197) = -0.37, p = .714$; $\mathrm{BF}_{01} = 11.78$), and they were similarly insensitive to the search asymmetry for Cs and Os ($M = 1.48$, 95% CI $[-2.87,5.82]$, $t(198) = 0.67, p = .503$; $\mathrm{BF}_{01} = 10.12$). The results with respect to Q and Os were more nuanced, with a marginally significant difference of 4.05 ms/item between O-in-Q and Q-in-O estimate slopes ($M = 4.05$, 95% CI $[-0.11, 8.21]$, $t(199) = 1.92$, $p = .056$; $\mathrm{BF}_{01} = 2.09$). However, even here, a difference of 4 ms/item in search time estimates is more than 10 times smaller than the true difference of 53.44 ms/item in slopes obtained from actual searches. Interestingly, asymmetries in the mean estimated search time (rather than the expected change per addition of one distractor) were somewhat stronger (orientation: mean difference of 21 ms, $t(197) = 1.79$, $p = .074$; open edges: mean difference of 21 ms, $t(198) = 1.74$, $p = .084$; line addition: mean difference of 40 ms, $t(199) = 3.66$, $p < .001$). However, here too, these effects are much smaller than the true effects in actual search behavior (mean differences of 250, 387, and 556 milliseconds for the three stimulus pairs, respectively).

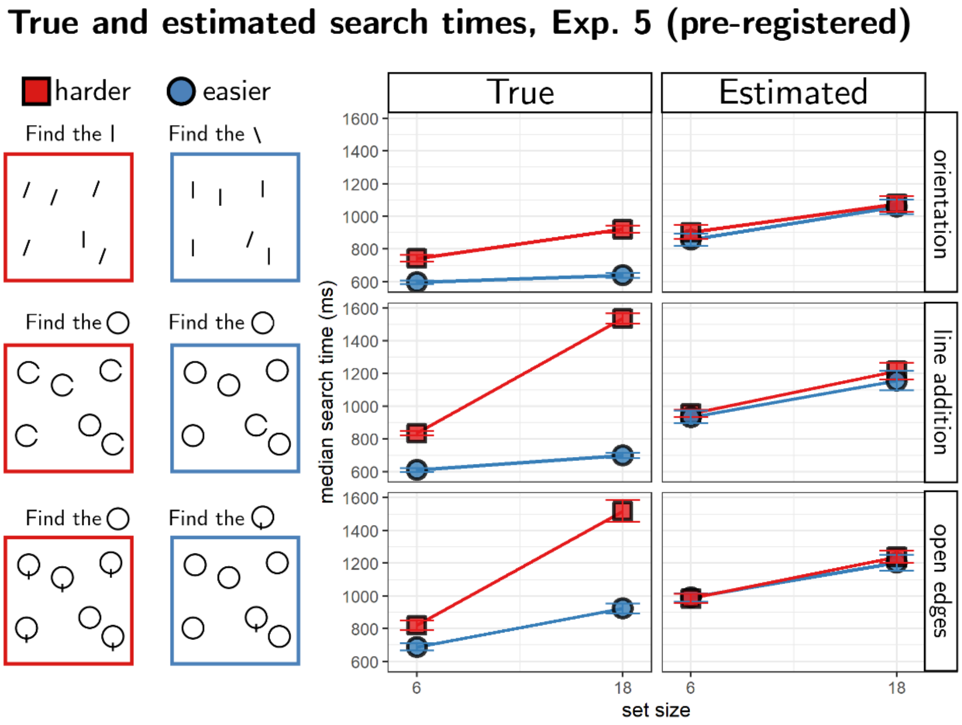


Figure 7. Results from Exp. 5. True and estimated median search times for the six different searches. Within each pair, the easier search (finding a tilted target among vertical distractors, a Q among Os, and a C among Os) appears in blue, and the reciprocal, harder, search in red.

Experiment 6: semantic versus visual similarity

Search difficulty is a function, among other things, of the similarity between the target and the distractors. If this fact is represented in internal models of visual search, the question remains what kinds of similarity affect people's intuitions about search difficulty, and whether they are the same ones that affect visual search in practice. To test this, in Exp. 6 we manipulated semantic and visual similarity between targets and distractors, and measured their independent effects on search times and search time estimates. Exp. 6 was pre-registered (pre-registration document: doi.org/10.17605/OSF.IO/AH9NR). Raw data, experiment demos, and full analysis scripts are available at github.com/matanmazor/metaVisualSearch.

Participants

For Exp. 6, 150 participants were recruited from the Prolific crowdsourcing web-service. The experiment took about 10 minutes to complete. Participants were paid $2. The highest performing 30% of participants received an additional bonus of $2.

Procedure

The procedure for Experiments 6 was similar to that of Exp. 5. Participants estimated their prospective search times for six different searches: three searches with the letter E and three with the number 3 serving as targets. Each target was used in three conditions involving distractors that could be semantically and visually similar to the target (baseline condition: the letters H and A for the target E or the numbers 8 and 2 for the target 3), semantically dissimilar but visually similar (the numbers 8 and 2 for the letter E and the letters A and H for the number 3), or semantically similar but visually dissimilar (same as the baseline condition, but appearing in a different font (italics versus not) relative to the target letter; see Fig. 8, left panel). For each target and condition, subjects provided estimates for set sizes of 6 and 18.

Results

Accuracy in the visual search task was high ($M = 0.95$, 95% CI [0.94,0.97]). Error trials and visual search trials that took shorter than 200 milliseconds or longer than 5 seconds were excluded from all further analysis. Participants were excluded if more than 30% of their trials were excluded based on the aforementioned criteria, leaving 146 participants for the main analysis of Experiment 6.

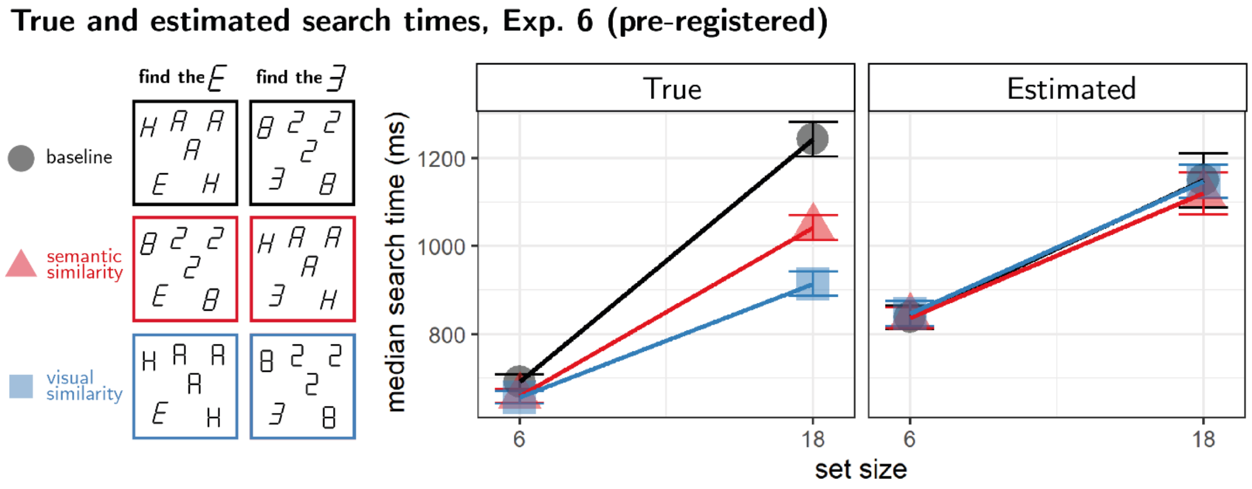


Figure 8. Exp. 6. Left: experimental conditions. Participants searched for an E or a 3 among semantically and visually similar (black circles, first row), semantically dissimilar but visually similar (red triangles, second row), or semantically similar but visually dissimilar distractors (blue squares, third row). Right: true and estimated search times for the three conditions.

Search slopes were significantly different for the six searches ($F(3.84,544.92) = 63.72$, $MSE = 0.00$, $p < .001$, $\hat{\eta}^2_G = .239$; see Fig. 8). Specifically, search slopes were significantly shallower when the target and distractors were of different semantic categories (31 ms/item) compared to the baseline condition (47 ms/item; $t(142) = -6.89$, $p < .001$). Visual dissimilarity produced an even stronger effect on estimation slopes (24 ms/item in the visual dissimilarity condition versus 47 ms/item in the baseline condition; $t(143) = -8.19, p < .001$). Finally, searching among number distractors was overall harder than searching among letter distractors ($t(142) = -15.22, p < .001$; See Fig. A2 for distractor-specific effects).

Prospective search time estimates were significantly correlated with actual search times (within-subject Spearman correlations: $M = 0.49$, 95% CI $[0.44{,}0.54]$, $t(145) = 20.72, p < .001$). Similar to search slopes, estimation slopes were also significantly different for the six searches ($F(4.11{,}583.14) = 3.92$, $MSE = 0.00, p = .003$, $\hat{\eta}_G^2 = .012$). However, unlike true search times, here we find no significant differences as a function of semantic dissimilarity (25 ms/item for both the semantic dissimilarity and baseline conditions; $t(142) = 0.42, p = .673$), or visual similarity (24 ms/item and 25 ms/item in the visual dissimilarity and baseline conditions; $t(143) = -1.11, p = .269$). Subjects were however sensitive to the main effect of distractor type on search time, producing steeper estimation slopes for numbers (26 ms/item) than for letters (22 ms/item; $t(142) = -2.96, p = .004$; See Fig. A2). Together, target-distractor similarity along the manipulated dimensions had significant effects on search difficulty, but we found no trace for these effects in search time estimates.

Discussion

Over more than four decades of research on spatial attention, experiments where participants report the presence or absence of a target in a display revealed basic principles such as the set-size effect (Treisman, 1986; Treisman & Sato, 1990; Wolfe, 1998), the advantage for feature search over more complicated conjunction and spatial configuration searches (Treisman, 1986; Treisman & Sato, 1990), and asymmetries in the representations of visual features (Malinowski & Hübner, 2001; Shen & Reingold, 2001; Treisman & Gormican, 1988; Treisman

& Souther, 1985). Some of these findings are intuitive, but others are more surprising; even without training in psychology, people have a set of expectations and beliefs about their own perception and attention, and about visual search more specifically.

Here we measured these expectations and their alignment with actual visual search behavior. In six experiments, we show that naive participants provide search time estimates that are consistent with partial metacognitive knowledge of their own attention. In line with previous reports, prospective search time estimates reflected accurate knowledge of the set size effect and differences in efficiency between feature and conjunction searches (Levin & Angelone, 2008; Miller & Bigi, 1977). Participants represented search efficiency along a continuum, were able to provide reasonably accurate search time estimates for complex stimuli and displays with which they had no prior experience, and were sensitive to the assignment of stimuli to target and distractor roles. At the same time, they showed little to no insight into visual search asymmetries and the effects of semantic target-distractor similarity on search efficiency, and as a group their estimates revealed no awareness of the pop-out effect for color search. In the following paragraphs, we unpack our central findings in more detail.

Do subjects know that color pops out?. Searching for a deviant color is relatively easy, and people know that. Psychology students correctly estimated that searching for a green vertical line is harder when some distractors are green compared to when all are red, and that increasing the number of distractors would make the search harder in the former, but not in the latter all-red case (Levin & Angelone, 2008). The understanding that adding more distractors does not affect search time in color search reflects metacognitive insight into the parallel nature of color search. Similarly, when asked to order visual search displays according to difficulty, 81% of third graders used color, but only 48% used shape, to inform their orderings (Miller &

Bigi, 1977). Knowledge about the special status of color is also evident in the way we communicate with others about what we see. People consistently prefer object descriptions that include information about color, even when color information is fully redundant (Jara-Ettinger & Rubio-Fernandez, 2021). To account for this fact, Jara-Ettinger and Rubio-Fernandez (2021) suggested that speakers hold mental models of the visual search behavior of their listeners, and choose their words to maximize search efficiency according to these models. In their hypothetical implementation of this model, color information allows listeners to restrict their search to objects of the target's color, making the search highly efficient. Thus, knowing that listeners can easily orient their attention by color, speakers prefer longer descriptions that reduce the effective set size and by that improve search efficiency.

In Experiments 1 and 2, we similarly found evidence for metacognitive knowledge that color search is easy. Estimation slopes for color search were consistently shallower than for orientation-color and shape-color conjunction searches (for comparison, estimation slopes for shape and orientation searches showed no such difference). Still, although shallower than conjunction estimation slopes, color estimation slopes were significantly positive at the group level, reflecting a belief that color search is serial in nature. This seems to be in conflict with the results of Levin and Angelone (2008), where only 32.5% of subjects thought that adding more distractors to a color search would make it slower (compared to 87.9% for color-orientation conjunction search). However, two differences between our studies are worth pointing out. First, Levin and Angelone's sample consisted of students, who may have learned or heard about visual search, and updated their internal models accordingly. And second, the fact that the mean estimation slopes in our experiments were overall positive does not preclude the possibility that for a subset of participants it was in fact zero. Using the proportion of positive estimation slopes for color search in Experiments 1 and 2 (0.71), and into the fact that this proportion should equal

0.5 among subjects who believe that set size has no effect on color search but their estimates are affected by random noise, we can extract a lower bound for the proportion of subjects who believed color search had a positive slope p by solving the equation $p + 0.5(1 - p) = 0.71$, resulting in an estimate of $p = 0.42$. Note that this analysis conservatively assumes that estimate slopes should be positive for all subjects who believed color search was serial. In other words, among our random sample of online participants, more than 40% of subjects provided estimates that are consistent with color search being serial.

This blindness to pop-out effects indicates a missing component in internal models of visual search (or at least, in the models of some participants): unlike Feature Integration Theory and Guided Search models, they have no pre-attentive components. This means that items are randomly selected in no particular order, and the only thing that changes between easier and harder searches is the speed with which this serial process can take place. Without a bottom-up activation of 'feature maps', or effortless processing of guiding signals, this model echos early theories of vision as a sense that operates more like touch than like hearing, by sending out sensors to explore the environment (for a review, see Dedes, 2005). The immediate pop-out of color cannot take place in a model that requires subjects to voluntarily attend to individual items in order to perceive them.

Beliefs about the relative efficiency of different search tasks can also be probed by measuring the time participants take to conclude that a target is absent from a display. Unlike target-present trials that are terminated upon detecting the target, in target-absent trials decisions are made based on the belief that a hypothetical target would have been found. For example, if subjects know that color search is parallel, they may immediately conclude that a target is absent from an array if the target color does not immediately pop out to their attention. In contrast, subjects that hold the erroneous belief that finding a color requires a serial search will take longer to conclude that a target is missing. Using this indirect approach, and focusing on the first trials of the experiment, before subjects have the opportunity to adapt their search termination strategies, Mazor and Fleming (2022) found that subjects immediately terminate a search when the target color is absent from the search array. This provides indirect evidence that the implicit metacognitive knowledge that is involved in guiding search termination is dissociable from the kind of explicit metacognitive knowledge that we measure here. In support of this dissociation between search termination behavior and explicit metacognitive ratings of search difficulty, Mazor and Fleming found that search termination slopes were shallower for feature searches than for conjunction searches even among participants whose explicit metacognitive ratings reflected a belief that feature searches are harder.

What is person-specific about internal models of visual search?. In Experiments 3 and 4, we show that internal models of visual search are at least partly person-specific: participants' predictions better fitted their own search times compared to the search times of other participants. Still, in both experiments, the correlation between participants' estimates and the search times of other participants was considerably above zero (see Fig. 5). We note that above-zero self-other correlations are expected even if internal models of visual search are fully person-specific, as long as search behavior is relatively conserved across different individuals. In contrast, a significant difference between self-self and self-other correlations is expected only if some of the knowledge that is expressed in search time estimates relies on idiosyncratic knowledge. We consider two possible sources of inter-subject variation that may contribute to idiosyncratic beliefs about visual search: judgments about similarity or complexity of visual objects, and person-specific knowledge about attention.

First, subjects may vary in how they perceive different visual objects to be simple or complex, similar or different. If perceptions of complexity and similarity contribute to search behavior, and if subjects' internal models correctly specify these effects of complexity and similarity on search behavior, generic internal models of visual search may produce person-specific search time estimates. Indeed, we found an advantage for self-self correlations only in Experiments 3 and 4, where stimuli were complex enough to produce meaningful variability in how they are perceived by different subjects. However, as we show in Exp. 4 and 5, any person-invariant specification of how similarity contributes to search time would need to be sensitive to asymmetries in the perception of similarity (Tversky, 1977) in order to fully account for our findings of a drop in the correlation between estimated and true search times when swapping the target and distractor roles. For example, internal models may specify that what matters most to search time is whether the target is similar to the distractors, but not so much whether the distractors are similar to the target.

Second, beliefs about attention itself may be learned or calibrated based on first-person experience. Humans accumulate observations not only of external events and objects, but also of their own cognitive and perceptual states. Specifically, subjects have been shown to notice when their attention is captured by a distractor (Adams & Gaspelin, 2021) even in the absence of an overt eye movement (Adams & Gaspelin, 2020). These observations can then be integrated into an internal model or an intuitive theory: which items are more or less likely to capture attention, under what circumstances, etc. Future research into the development of this simplified model and its expansion based on new evidence (for example, by measuring intuitions before and after exposure to some evidence, Bonawitz, Ullman, Bridgers, Gopnik, & Tenenbaum, 2019) is needed to understand the relation between metacognitive monitoring of attention and metacognitive knowledge of attentional processes.

This relates to recent theoretical and empirical advances underscoring the utility of keeping a *mental self-model*, or a *self-schema* for attention control (Wilterson et al., 2020), social cognition (Graziano, 2013) , phenomenal experience (Metzinger, 2003), and inference about absence (Mazor, 2021; Mazor & Fleming, 2022). For example, knowing that a red berry would be easy to find among green leaves, a forager can quickly decide that a certain bush bears no ripe fruit. Alternatively, knowing that a snake would be difficult to spot in the sand, they might allocate more attentional resources to scanning the ground. Reasonably accurate search time estimates in Experiments 3 and 4 suggest that internal models of spatial attention that can be applied to unseen stimuli in novel displays, and are at least partly tailored to one's own perceptual and cognitive machinery.

What is the role of target-distractor similarity?. Visual search is harder when distractors are similar to the target. Having insight into this simple fact, and the fact that searches become harder with the addition of more distractors, should be sufficient to produce search time estimates that are aligned with actual search times. This way, subjects can rely on a rough overlap between items that are similar to the target and ones that have the potential to be distracting, and produce relatively accurate search time estimates based on their similarity judgments alone. Alternatively, as described above, subjects may be using an approximate probabilistic model of their own attention, producing search time estimates that correlate with, but are not causally dependent on, perceptions of similarity.

We set up Exp. 6 to directly test the effect of target-distractor similarity on perceived search difficulty. To our surprise, search time estimates were not at all sensitive to sizeable effects of semantic and visual target-distractor similarity on search time. This metacognitive blindness was specific to the interaction between target and distractor identities: subjects were sensitive to the fact that number distractors are more distracting overall – an effect that we, based on our knowledge of scientific models of visual search, could not predict in advance.

Importantly, this finding is consistent with search time estimates being causally dependent on perceptions of similarity, just not the kind in which a number is similar to a number more than to a letter, or a tilted character more similar to tilted than to upright characters. Instead, in this hypothetical similarity space number distractors in Exp. 6 were more similar to both 3s and Es than were letter distractors. More broadly, if search time estimates are driven by implicit judgments of target-distractor similarity, they seem to be selective to specific similarity metrics that are perceived as being relevant to search behaviour. As a result, even this metacognitively lean account of our findings requires subjects to hold nuanced beliefs about their visual search behaviour. As we discuss above, internal models of visual search may alternatively produce search time estimates based on an approximate probabilistic model, and without any reference to target-distractor similarity. More work is needed to determine where internal models of visual search fall on the continuum between being simulation-based and being rule-based, noting that much of human knowledge may lie somewhere in between these two ends (Bass, Smith, Bonawitz, & Ullman, 2021; Hegarty, 2004).

Conclusion. Across six experiments, we observed the following patterns in prospective search time estimates: First, estimates correlated with search times (Exp. 1-6). Second, estimates reflected knowledge of the set-size effect (Exp. 1, 2, 5 and 6), and were biased to assume a set-size effect even in searches that are in fact parallel (Exp. 1, 2). Third, estimates were sensitive to the relative efficiency of different searches (Exp. 1-6), even for complex, unfamiliar stimuli (Exp. 3, 4). Fourth, estimates were sensitive to the assignment of stimuli to target and distractor roles (Exp. 4, 5), but did not show reliable search asymmetries for basic visual features (Exp. 5). Finally, estimates were not sensitive to direct manipulation of target-distractor similarity along visual and semantic dimensions (Exp. 6).

Together, our results reveal that search difficulty is represented along a gradient in a person-specific manner, but that this representation is limited. Most notably, we find that subjects have no metacognitive access to pre-attentive stages of visual search. Our findings place a lower bound on the richness and complexity of subjects' internal model of visual search and attention more generally, opening a promising avenue for studying humans' intuitive understanding of their own mental processes.

References

Adams, O. J., & Gaspelin, N. (2020). Assessing introspective awareness of attention capture. *Attention, Perception, & Psychophysics*, 1–13.

Adams, O. J., & Gaspelin, N. (2021). Introspective awareness of oculomotor attentional capture. *Journal of Experimental Psychology: Human Perception and Performance*.

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*.

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, *38*(7-8), 413–424.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Blakemore, S.-J., Wolpert, D. M., & Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nature Neuroscience*, *1*(7), 635–640.

Bonawitz, E., Ullman, T. D., Bridgers, S., Gopnik, A., & Tenenbaum, J. B. (2019). Sticking to the evidence? A behavioral and computational case study of micro-theory change in the domain of magnetism. *Cognitive Science*, *43*(8), e12765.

Brown, J., Lewis, V., & Monk, A. (1977). Memorability, word frequency and negative recognition. *The Quarterly Journal of Experimental Psychology*, *29*(3), 461–473.

Dedes, C. (2005). The mechanism of vision: Conceptual similarities between historical models and children's representations. *Science & Education*, *14*(7), 699–712.

Forrester, J. W. (1971). Counterintuitive behavior of social systems. *Theory and Decision*, *2*(2), 109–140.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. *Oxford Handbook of Causal Reasoning*, 515–548.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Mit Press.

Graziano, M. S. (2013). *Consciousness and the social brain*. Oxford University Press.

Graziano, M. S., & Webb, T. W. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology*, *6*, 500.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Henninger, F., Shevchenko, Y., Mertens, U., Kieslich, P. J., & Hilbig, B. E. (2019). *Lab. Js: A free, open, online study builder*.

Jara-Ettinger, J., & Rubio-Fernandez, P. (2021). The social basis of referential communication: Speakers construct physical reference based on listeners' expected visual search. *Psychological Review*.

Lake, B., Salakhutdinov, R., Gross, J., & Tenenbaum, J. (2011). One shot learning of simple visual concepts. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*.

Lange, K., Kühn, S., & Filevich, E. (2015). " just another tool for online studies"(JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS One*, *10*(6).

Levin, D. T., & Angelone, B. L. (2008). The visual metacognition questionnaire: A measure of intuitions about vision. *The American Journal of Psychology*, 451–472.

Malinowski, P., & Hübner, R. (2001). The effect of familiarity on visual-search performance: Evidence for learned basic features. *Perception & Psychophysics*, *63*(3), 458–463.

Mazor, M. (2021). *Inference about absence as a window into the mental self-model.*

Mazor, M., & Fleming, S. M. (2022). Efficient search termination without task experience. *Journal of Experimental Psychology: General*.

Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, *2*(4), 353–393.

Miller, P. H., & Bigi, L. (1977). Children's understanding of how stimulus dimensions affect performance. *Child Development*, 1712–1715.

Shen, J., & Reingold, E. M. (2001). Visual search asymmetry: The influence of stimulus familiarity and low-level features. *Perception & Psychophysics*, *63*(3), 464–475.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Siegel, M. H., Magid, R. W., Pelz, M., Tenenbaum, J. B., & Schulz, L. E. (2021). Children's exploratory play tracks the discriminability of hypotheses. *Nature Communications*, *12*(1), 1–9.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Treisman, A. (1986). Features and objects in visual processing. *Scientific American*, *255*(5), 114B–125.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, *95*(1), 15.

Treisman, A., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 459.

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

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

Wilterson, A. I., Kemper, C. M., Kim, N., Webb, T. W., Reblando, A. M., & Graziano, M. S. (2020). Attention control and the attention schema theory of consciousness. *Progress in Neurobiology*, *195*, 101844.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science*, *9*(1), 33–39.

Wolfe, J. M. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, 1–33.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 419.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*(5232), 1880–1882.

Zhang, Y. R., & Onyper, S. (2020). Visual search asymmetry depends on target-distractor feature similarity: Is the asymmetry simply a result of distractor rejection speed? *Attention, Perception, & Psychophysics*, *82*(1), 80–97.

Appendix

Incentive structure

We assume that participants represent the distribution of response times conditional on a specific search array as a right-skewed, positive distribution. Here, we assume that internal distributions of response times abide by the rule that

$$log(RT) \sim N(\mu, \sigma)$$

where σ is fixed per participant, and μ varies as a function of search difficulty.

In the estimation part, participants produces an estimate x. In case they responds as fast, or faster, than their original estimate, they get a bonus of $\frac{10}{\sqrt{x}} = 10 \cdot x^{1/2}$. Since 10 is a constant, we ignore it in the following derivations. The expected bonus given for a trial is now:

$$E[bonus|x] = Pr_{logRT \sim N(\mu,\sigma)}[\log(x) \geq logRT] \cdot x^{-1/2}.$$

Which can be re-written for $log(x)$ as:

$$E[bonus|\log(x)] = Pr_{logRT \sim N(\mu,\sigma)}[\log(x) \geq logRT] \cdot e^{-\log(x)/2}.$$

Since we assumed log RTs are distributed normally, we can express $log(x)$ relative to the distribution of log RTs as $log(x) = \mu + \alpha \cdot \sigma$ for some number α. This number represents the position of the estimate relative to the distribution of response times, with lower values corresponding to more risky estimates, and higher values to more conservative ones. The expected bonus is then:

$$\begin{aligned} E[bonus|\alpha] &= Pr_{z\sim N(0,1)}[\alpha > z] \cdot e^{-(\mu+\alpha\cdot\sigma)/2} \\ &= Pr_{z\sim N(0,1)}[\alpha > z] \cdot e^{-(\alpha\cdot\sigma)/2} \cdot e^{-\mu/2}. \end{aligned}$$

where z is the standardized $log(RT)$, with mean 0 and standard deviation 1. μ only appears in the third term in the product, which functions as a constant multiplier that scales the expected bonus equally for all choices of α. It then follows that the function relating the choice of α to the expected bonus preserves its shape for all possible values of μ. This function reaches a maximum for $\alpha = 0.52$ (the 70th quantile) in Exp. 1-4, and for $\alpha = 0$ (the 50th quantile) in Exp. 5:

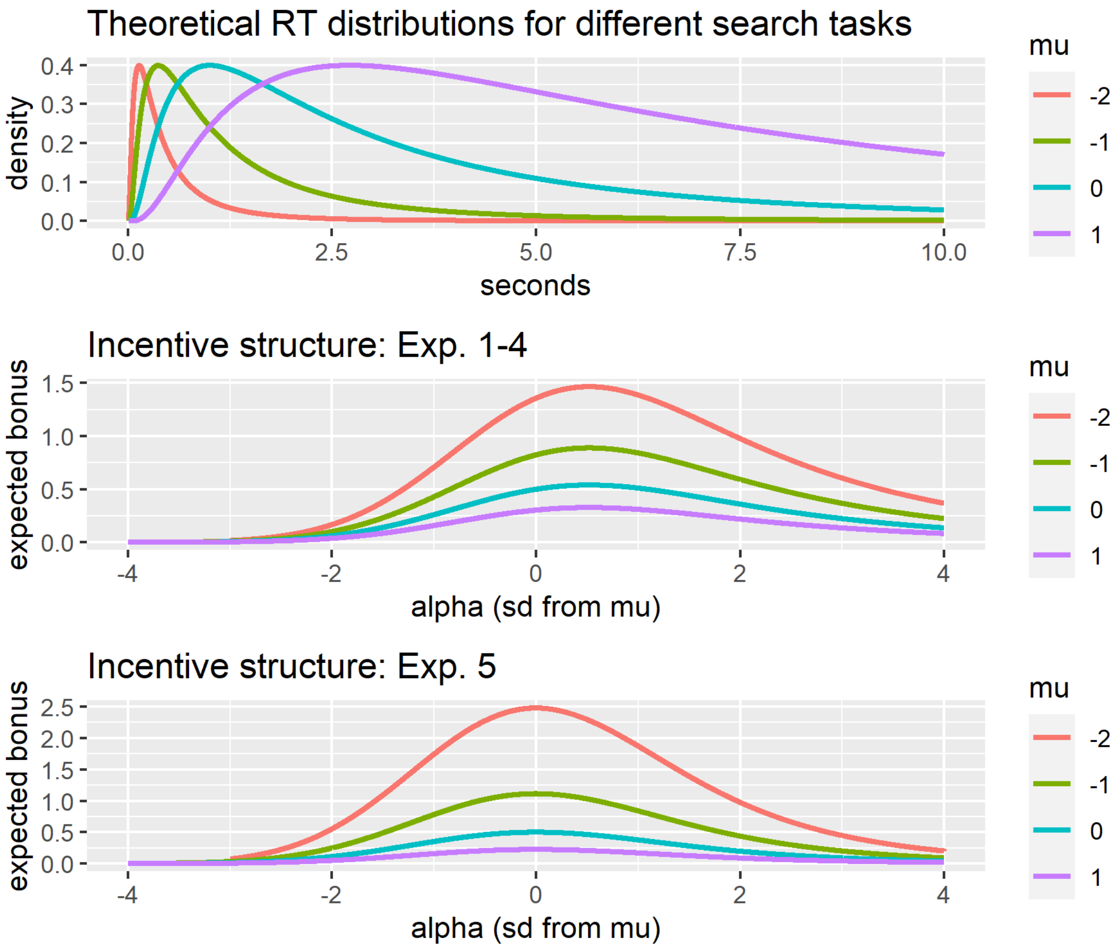


Figure A1. Upper panel: response time distributions are modeled as exponents of values drawn from a normal distribution with different values of mu. Middle panel: the estimate value that maximizes the expected bonus is fixed with respect to the mean of the log(RT) distributions, regardless of what the mean is. The expected bonus is higher for lower values of mu, but to maximize their bonus participants should always choose an estimate that is positioned in the 70 quantile of the RT distribution (mean + 0.518 standard deviations in log space). Lower panel: in Exp. 5, the bonus is maximized for estimates that are aligned with the median of the RT distribution.

Exp. 6 additional analyses

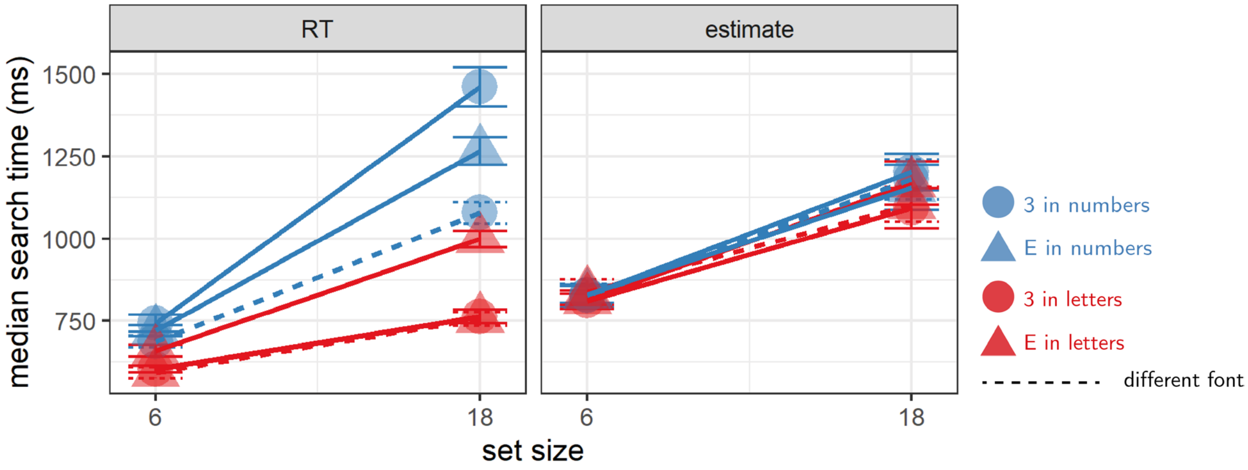


Figure A2. Search times and search time estimates as a function of target and distractors, Exp. 6. Error bars represent the standard error of the median, estimated with bootstrapping.