Atten Percept Psychophys (2016) 78:2135–2151
DOI 10.3758/s13414-016-1128-1

When is it time to move to the next map? Optimal foraging in guided visual search

Krista A. Ehinger[1] · **Jeremy M. Wolfe**[1]

Published online: 18 May 2016

Abstract Suppose that you are looking for visual targets in a set of images, each containing an unknown number of targets. How do you perform that search, and how do you decide when to move from the current image to the next? Optimal foraging theory predicts that foragers should leave the current image when the expected value from staying falls below the expected value from leaving. Here, we describe how to apply these models to more complex tasks, like search for objects in natural scenes where people have prior beliefs about the number and locations of targets in each image, and search is guided by target features and scene context. We model these factors in a guided search task and predict the optimal time to quit search. The data come from a satellite image search task. Participants searched for small gas stations in large satellite images. We model quitting times with a Bayesian model that incorporates prior beliefs about the number of targets in each map, average search efficiency (guidance), and actual search history in the image. Clicks deploying local magnification were used as surrogates for deployments of attention and, thus, for time. Leaving times (measured in mouse clicks) were well-predicted by the model. People terminated search when their expected rate of target collection fell to the average rate for the task. Apparently, people follow a rate-optimizing strategy in this task and use both their prior knowledge and search history in the image to decide when to quit searching.

Keywords Visual search · Foraging · Search termination · Satellite imagery · Guided search · Absent trials

✉ Krista A. Ehinger
k.a.ehinger@gmail.com

[1] Visual Attention Lab, Harvard Medical School, Brigham & Women's Hospital, 64 Sidney St. Suite 170, Cambridge 02139, MA, USA

In a classic visual search task in the laboratory, an observer looks for a target item among some number of distractor items. The single target is present or absent, and a search ends when the target is found or the observer abandons the search, declaring the target to be absent. This all occurs over the course of a few hundred to a few thousand milliseconds. A great deal is known about such searches (for some recent reviews, see Chan & Hayward, 2013; Eckstein, 2011; Wolfe, 2014; Wolfe, Horowitz, & Palmer, 2010). For instance, we know that the efficiency of these searches falls on a continuum, as indexed by the slope of the function relating RT to set size (the number of items on-screen; Wolfe, 1998). The relationship of target to distractor items is a powerful determinant of search efficiency (Duncan & Humphreys, 1989). If the target differs from a homogeneous set of distractors on the basis of a basic attribute like color or motion, search will be extremely efficient. Indeed, the target will "pop-out" independent of the number of distractors (Egeth, Jonides, & Wall, 1972). If the target and distractors share all their features, differing only in their arrangement, search will be quite inefficient, even if the items are clearly resolvable in peripheral vision (Bergen & Julesz, 1983), perhaps reflecting serial deployment of attention from item to item (Kwak, Dagenbach, & Egeth, 1991). If a basic feature of the target can give partial information, attention will be *guided* by that information. For example, if the target, when present, is green, and only half the distractors are green, then attention will be guided to green items (Egeth, Virzi, & Garbart, 1984), and the efficiency will be double what it would have been without the color information. Hence, the idea of "guided search" (Wolfe 1994, 2007, 1989) with a limited set of attributes available to guide (Wolfe and Horowitz, 2004).

This body of research will tell you something about searching for your car in the parking lot (if it is a red Prius, don't waste time attending to blue cars) or the bottle opener in

the kitchen drawer (this will be inefficient due to a lack of a salient defining feature, not to mention "crowding" effects;(Balas, Nakano, & Rosenholtz, 2009). But let us consider a different search task. Suppose you are tasked with searching for military vehicles in satellite images of the tense border between two countries. This search differs in a variety of important ways from classic, laboratory search.

1) In a continuous scene, it is going to prove essentially impossible to measure the set size (Neider & Zelinsky, 2008; Wolfe, Alvarez, Rosenholtz, & Kuzmova, 2011), rendering the idea of search efficiency problematic.
2) Even if we could count the items, we do not know how many items are processed in a single fixation. The eyes move 3–5 times per second, and it is tempting to assume that each fixation catalogs a target/nontarget decision about a single item. However, in simple searches, at least, items are processed at much higher rate. Multiple items might be processed in parallel during each fixation. Serial attention might visit multiple items on each fixation. Indeed, both parallel processing and serial selection probably characterize search (Wolfe, 2003).
3) The number of targets is unknown, meaning that, even if you find a target, you still do not know, for certain, that it is time to end the search of this image.
4) The search is guided, but not merely by basic attributes of the target. The structure of the scene tells the searcher where targets are more or less likely (very few vehicles in trackless wilderness or in the middle of lakes; Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009; Torralba, Oliva, Castelhano, & Henderson, 2006; Võ & Wolfe, 2015).
5) The tasks have a more striking learning component. People are trained to do these complex tasks, and one manifestation of that learning is that experts learn where *not* to look (Kundel & La Follette, 1972; Wooding, Roberts, & Phillips-Hughes, 1999).
6) Finally, each stimulus here will be searched for minutes rather than for a fraction of a second.

The purpose of this paper is to ask how people perform tasks like the one described here; tasks that can be called "extended search" tasks. These tasks include search in overhead imagery as well as other prolonged search tasks, such as search for signs of disease in radiology or search for cracks in the examination of an airplane. A major difference between these tasks and more classic, fast visual search tasks is the role of higher level knowledge and decision-making processes. Extended search tasks are slow and can involve rich, complex natural images, so the observer's top-down knowledge and expectations about the images and task are an important component of their search in these images.

With difficult, multiple target search in a complex scene, the decision to end search of the current image becomes one of the most interesting problems. After all, since it is hard to find everything, one could keep looking for a long time, but there are many more images to search. When have the diminishing returns diminished to the point where it is time to move on?

The topic of search termination has been studied in human visual searches having zero or one target (Chun & Wolfe, 1996; Cousineau & Shiffrin, 2004; Moran, Zehetleitner, Müller, & Usher, 2013; Wolfe, 2012). Quitting times in searches with multiple targets have been studied extensively in the animal foraging literature, where it is described as the "patch-leaving" problem (Stephens, Brown, & Ydenberg, 2007; Stephens & Krebs, 1986). If you are grazing in this spot or sipping nectar from flowers on this plant, when should you leave for the next patch of grass or the next flowering plant? More recently these rules have been applied to human searches for visual stimuli (Wolfe, 2013), information (Pirolli & Card, 1999), or even the contents of one's own memory (Hills, Todd, & Jones, 2015). One of the earliest and most influential quitting-time models in the animal literature is the marginal value theorem (MVT) proposed by Charnov (1976). This theorem considers the problem of an animal foraging for food in an environment where food is randomly distributed in many separate patches, and assumes that the animal's goal is to maximize its rate of food intake. While feeding in a patch, the animal gradually exhausts the food supply, and the intake rate in that patch drops. However, traveling to a new patch imposes a cost: It takes time, and no food can be collected while traveling between patches. The optimal strategy is to leave the patch when the expected rate from traveling to a new patch exceeds the expected rate from staying in the current one. According to MVT, this occurs when the rate of food intake in the current patch falls to the average rate for the environment (see Fig. 1).

This theory is appealing because it claims that the optimal patch-leaving time can be computed from a single, easily observed variable: the current rate of food intake in the patch. There are some models of search for which it makes sense to assume a continuous rate. For example, limited capacity and decision integration models of search propose that the whole visual field is processed in parallel, and target detection is actually a signal detection problem across various locations in the visual field (Palmer, 1995; Palmer, Verghese, & Pavel, 2000; Townsend, 1971;). The gradual accumulation of information across the visual field could be represented as a smooth, continuous "intake rate" curve. However, computing this rate is not so straightforward for tasks that involve slow, serial search for discrete targets. When a forager is collecting individual items (pieces of fruit, prey animals, tumors, military vehicles, etc.), the intake curve looks more like the step function in Fig. 1 (dotted line): The intake rate is zero while the observer is searching, then jumps sharply when the searcher finds a target. It wouldn't make sense for the searcher to leave the patch shortly after the instantaneous rate falls below the average – the instantaneous rate in a patch can be zero for a

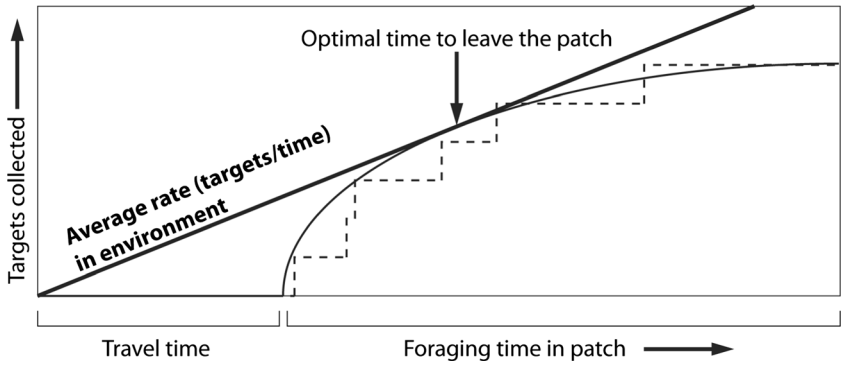

Fig. 1 Illustration of optimal foraging theory. The solid line represents an idealized forager's intake over time. During the "travel time" period, the forager is moving to the patch, and intake is zero. Once arriving in the patch ("foraging time"), intake increases rapidly at first, then gradually declines as the patch is exhausted. If the goal is to maximize rate of target collection, the optimal time to leave the patch is when the intake rate falls to the average in the environment (bold diagonal line); this is the point where the intake curve is tangent to the average rate. But a forager doesn't actually experience this solid-line curve while foraging in a single patch; instead, it collect discrete targets at random intervals, as represented by the dotted line. An optimal forager must infer the expected rate (solid line) and optimal leaving time from its experienced rate (dotted line) and expectations about the patch

significant time between target detections. As a proxy, the searcher might compute the time since it last found a target and use that as a measure of the current rate; it would leave the patch whenever the time since last target exceeded a threshold determined by the average rate in the environment. This "giving up time" (GUT) rule does seem to explain foraging behavior in certain situations (Krebs, Ryan, & Charnov 1974).

However, there are serious problems with the giving up time rule and other similar implementations of the marginal value theorem. In most searches, the time between targets is somewhat noisy: It's randomly distributed around an expected value. An ideal forager should make decisions based on the expected rate, not the experienced rate. Also, a forager might be expected to learn about current patch quality from the success or failure of its ongoing search in the current patch: A patch where many targets are found quickly is probably a rich patch and might be worth spending more time in. To some extent this is captured by MVT (the rich patch should have a higher instantaneous rate, so a forager should stay there longer), but it's not clear that the giving up time is actually an optimal leaving strategy. In fact, simulations show that it is not: In environments where patch quality varies and the experienced rate of target collection is noisy, the giving up time strategy proposed by MVT is not optimal and can be made to perform arbitrarily badly, depending on the amount of variation in the environment (Green, 1980, 1984; McNamara, 1982; Oaten, 1977;).

As Oaten (1977) points out, MVT is flawed because it assumes that stochastic aspects of the foraging task balance out and the forager can make good decisions based on averages. In fact, an optimal forager should reason about the foraging task probabilistically. Various frameworks for this have been proposed, including the original stochastic foraging model of Oaten (1977), but Bayesian optimal foraging models (Green, 1980; McNamara, Green, & Olsson, 2006; McNamara & Houston, 1985) are probably the easiest to generalize. According to these approaches, leaving decisions are made based on the *potential* value of the patch: The optimal leaving time is when the expected rate, not the observed rate, drops below the average for the environment (McNamara, 1982). By "expected rate" we mean an estimate based on the forager's belief about how many targets are in the patch and how easy they should be to find. These beliefs are updated in a Bayesian fashion as the forager searches for targets. For example, suppose that you come across a garage sale or moving sale where someone has an assortment of household items displayed. On first glance, it looks quite uninteresting, so you decide to forage briefly because your expected rate of return is low. However, if you find a surprising treasure, the expected value goes up, and you should search longer.

The difference between the MVT and potential value approaches can be illustrated using Fig. 1. In this graph, time in a patch is shown on the *x*-axis and targets collected on the *y*-axis, and the dotted line shows an individual forager collecting six targets in the patch. Since targets are discrete objects, they appear as steps: The width of the step indicates the time elapsed between collecting one target and the next. The solid straight line represents this forager's average rate of target collection, and according to MVT, the forager should quit searching the patch when its instantaneous rate falls below this average rate. The instantaneous rate is one over the time elapsed since collecting the last target, or the slope drawn from the corner of the current step to the last one. An MVT forager would use this slope to decide when to leave the patch. On the other hand, a potential value forager would try to model the average expected rate of target collection, shown by the solid curved line. Potential value foragers would use the same leaving time threshold – they would quit when their current rate fell below the average for the task – but they would use the slope of their expected rate (solid curve) instead of their experienced rate (dotted line) to decide if their current rate was below that threshold.

Target-present/target-absent search is another example that can illustrate the difference between the MVT and potential value approaches (discussed in more detail in McNamara & Houston, 1985). In this kind of task, there can only be one target per patch, but not all patches have a target, so the forager must either find the target or give up the patch as empty and move on (for simplicity, we'll assume that search is random with replacement: The forager can't prioritize the locations most likely to have a target or search the entire patch exhaustively). Bayesian foragers would start with some initial beliefs about how likely the patch was to have a target and update that belief while searching. After every unsuccessful search attempt, they would be slightly more inclined to think the patch was empty. The optimal forager would leave the patch immediately after finding a target (because at that point the patch is guaranteed to be empty and its potential value has become 0), or when the likelihood that the patch was empty was so high that the expected rate from staying was lower than the expected rate from leaving (exactly where this threshold occurs depends on the travel time cost of leaving a patch and traveling to the next one). A forager using a simple marginal value theorem approach, on the other hand, would quit the patch when the time since last finding a target exceeded the mean rate in the environment. Since the simple marginal value account would not include the understanding that a present/absent search is over when the observer finds the target, this would lead to the clearly incorrect prediction that the average target-present trial will actually be slower than the average target-absent trial, since even after finding the target foragers must continue searching the empty patch for a length of time equal to the mean time between targets in the environment in order to convince themselves that the patch is empty.

In the potential value framework, determining the optimal leaving time in a foraging task requires modeling the forager's mental representation of the task. Leaving time is going to depend on the potential value of a patch, but potential value can't be measured directly. It is based on the forager's current expectations. In turn, those expectations are based on the forager's prior beliefs about patch quality, the history of search in the patch, and the actual probability of finding or not finding a target on each search attempt. This is easy enough to compute for simple cases (e.g., if the patch contains a known number of "items" and these are searched in a random-with-replacement fashion), and most of the foraging literature has focused on these types of cases (e.g., Green, 1984; Oaten, 1977). Cain, Vul, Clark, and Mitroff (2012) extended this analysis to the case where there could be a small number of targets present in any display. The specific number in a display was drawn from a distribution of possible, small numbers of targets and observers changed their behavior in response to the manipulation of that distribution. However, as noted above, real-world search tasks, such as search for objects in scenes, are more complex. We can't count the number of items in the scene nor do we know how many items are processed in a single fixation. We can be reasonably sure that people start the search with some prior knowledge on how likely the target is to be in the scene. Their search will be guided strategically to probable locations and/or to things that share basic features with the target (Ehinger, et al., 2009; Wolfe, 2007).

In this paper, we use data from a novel satellite imagery search task to motivate a general search termination model that can be applied to both guided and random search tasks. We assume that people quit searching when their expected rate of target collection drops below a threshold. That expected rate is derived from prior beliefs about how many targets are likely to be in an image and the observer's sense of how difficult, on average, the search task should be. We also assume that people update their expectations during search in a Bayesian fashion, decreasing their expectations if the rate of target collection is slower than expected, or increasing their expectations if they find targets more quickly than expected.

The task

In our satellite imagery search task, the targets are gas stations. Gas stations were chosen because they are recognizable from above with a little training, most people have some sense of where gas stations are most likely to appear in images (e.g., near major roads), and ground truth labels to denote which buildings are or are not gas stations are reasonably easy to obtain. At the scale at which we are presenting images, the gas stations are very small, so participants search each image using a magnifier interface, shown in Fig. 2. Since a gas station cannot be positively identified without the magnifier, we can use the magnifier as a rather literal "spotlight of attention" (Posner, Snyder, & Davidson, 1980); most conveniently, a spotlight that we can track. The magnifier allows us to record the locations searched in each map and makes it easier to fit the task into a Bayesian optimal foraging framework because we know what proportion of the map is captured in each magnified view. Search for gas stations in these images is a complex, guided search task similar to search in real-world scenes: Participants have prior beliefs about the quality of each map or "patch" (e.g., urban views should have more gas stations than rural ones), and they can search strategically, using their knowledge of what gas stations look like and where they are most likely to occur (e.g., more likely at intersections and unlikely in open fields).

In addition to investigating leaving times, we investigated whether different magnification interfaces have any effect on search. We compared an interface in which the magnified view was shown to the side of the overview map ("side-by-side") to interfaces in which the magnified view appeared in the map, either overlapping the zoom location ("magnifying glass") or just beside it ("offset"). There has been some previous work looking at how different magnification interfaces

Fig. 2 Interface and stimuli for the search task. (**a**), (**b**), and (**c**) show the three magnification interfaces used for the task. (d) shows a section of the overview map, slightly zoomed in for illustration purposes. There are two gas stations in this section of the map. Map imagery © Google, Digital Globe

affect visual search. Zhao, Rau, Zhang, and Salvendy (2009) found that people were able to complete a word search task more quickly when the magnified view appeared at the magnified location in the word search display (equivalent to our "magnifying glass" condition) than when it appeared off to the side of the display (as in our "side-by-side" condition). However, it is not clear whether these results extend to other types of visual search tasks. To anticipate our results, in this experiment, the type of magnifier did not have a significant effect on the results.

Experiment 1

Method

Participants

Sixty-two people participated in a web-based experiment on Amazon Mechanical Turk. Participants were based in the United States and had a good track record on the Mechanical Turk site (at least 100 HITs completed and an acceptance rate of at least 95%). Participants gave informed consent before starting the task. Payment was performance based: Participants received a base payment of $1 if they found at least 10 gas stations and a bonus of $0.10 per gas station for every gas station after the first 10.

Stimuli

The stimuli were 50 satellite view images from Google Maps. The overview image was 1,000 pixels square with a zoom level of 16, which corresponds to a real-world area of about 1.15 square kilometers. Views were chosen from 10 U.S. cities (five views per city). The magnified view was 200 pixels on each side; within this window the zoom level could be increased from level 16 to 19 (8× magnification).

The full overview images contained 0–10 gas station targets. Gas stations were identified by searching for "gas station" on each map and then using Google Maps' Streetview imagery to verify each result. We also manually searched each image to identify gas stations which appeared in the map but weren't included in the Google search results.

Design and procedure

Participants were randomly assigned to one of the three magnifier conditions: side-by-side, magnifying glass, and offset magnifying glass. At the start of the experiment, participants were asked to fill out a short demographic survey and were given instructions on the magnifier interface and shown examples of gas station targets. Participants were told the maximum number of trials (50) but were not given information about the number of targets per trial. The 50 overview images were shown in random order.

On each trial, participants were shown one overview image and asked to find the gas stations. The interface for a single trial in each condition is shown in Fig. 2. Participants could zoom into a part of the image by left-clicking on it. The + and – keys on the keyboard were used to increase and decrease the zoom level within the zoom through four possible zoom levels: 1×, 2×, 4×, and 8× magnification. The magnifier views could be closed by right-clicking within the zoom window. To mark a gas station, participants would center it in the zoom window and press the X key on their keyboard. They were then asked to rate their confidence that the building was a gas station on a scale from 1–9. The location was then marked, and participants were given feedback on their choice – a green marker meant that the location was a gas station; a red marker meant that it was not. Whenever the participant correctly marked a gas station, 10 points were added to a score total shown beside the map; this was included so participants could keep track of how many stations they had found so far and how much they would be paid. The map images were served from Google Maps and operations such as zooming, recentering, and marking were handled by Google Maps API.

Participants pressed a button to end each trial. At this point, participants were shown any gas stations they had missed in the current view. The feedback was intended to help participants learn what the targets look like and to give them an accurate count of the number of gas stations in each previously seen map so they could learn how many targets to expect in an average map in this task. After each trial, participants had the option to proceed to the next trial, or quit the task for now, which would pause the experiment clock. Participants could resume the task at a later time or, if they had found the minimum required number of targets, quit the task entirely, submit their work, and receive their bonus. Participants were not required to complete all 50 maps before quitting. There were no time limits in the trials, but participants were required to finish the task within 72 hours of starting it.

"*How many gas stations in this satellite image?*" *task*

To estimate observers' initial expectations for the number of gas stations in each satellite view, we ran a second Mechanical Turk task in which we showed each image at the lowest zoom level and asked 10 workers to guess the number of gas stations in the view without actually searching with the magnifier. Workers were told that it was possible for an image to have no gas stations but were not given any other range information. Workers were paid $0.01 per image and a $0.02 bonus for correct guess; they did not receive any feedback about their guesses during the task. The worker requirements and consent process for this task were the same as for Experiment 1. The averages of the guesses were moderately well correlated with the true numbers of targets in these images (by-images correlation: $r = 0.55$), which confirms that untrained participants have reasonable intuitions about the distribution of gas stations in these satellite images.

Results and discussion

We dropped 15 trials over 60 minutes in length (one supposes the observers went elsewhere, leaving the program running), 86 trials with no clicks recorded, and four trials which had recorded click locations incorrectly, leaving 1,541 trials. The percentage of trials dropped was slightly higher in the side-by-side condition (8% vs.5% in the other two magnifier conditions). Some of this difference is due to a single participant in the side-by-side condition who had 23 out of 50 trials dropped and was dropped entirely from the remaining analyses. The 61 Mechanical Turk participants contributed 5–50 trials each (mean 24, median 22).

We compared search speed and accuracy across the three magnifier conditions in a by-subjects analysis, shown in Fig. 3. Because of the random assignment of conditions and the fact that participants were not required to complete all the trials, there were an unequal number of participants and trials across the three magnifying conditions: 23 participants (535 trials) in the side-by-side condition, 17 participants (487 trials) in the magnifying glass condition, and 21 participants (492 trials) in the offset magnifying glass condition. Our measures of accuracy across conditions were precision (the proportion of targets found in each trial), recall (the proportion of participants "target" marks that were correct), and F score (an overall measure of accuracy equal to (2 * precision * recall)/(precision + recall)). Measures were computed within a trial then averaged across trials for each participant. There was no significant difference in the total time per map, $F(2, 58) = 1.5$, $p = .23$, number of search clicks per map, $F(2, 58) = 0.32$, $p = .72$, rate of clicks, $F(2, 58) = 0.13$, $p = .88$, average recall, $F(2, 58) = 2.28$, $p = .11$, average precision, $F(2, 58) = 0.27$, $p = .76$, or F score, $F(2, 58) = 0.09$, $p = .92$. The significance values did not change when only the participants who did more trials (at least 10 or at least 20) were included in the analysis. Since the three magnifier conditions did not seem to have any significant effect on search in these maps, we collapsed across magnifier conditions for all of the following analyses.

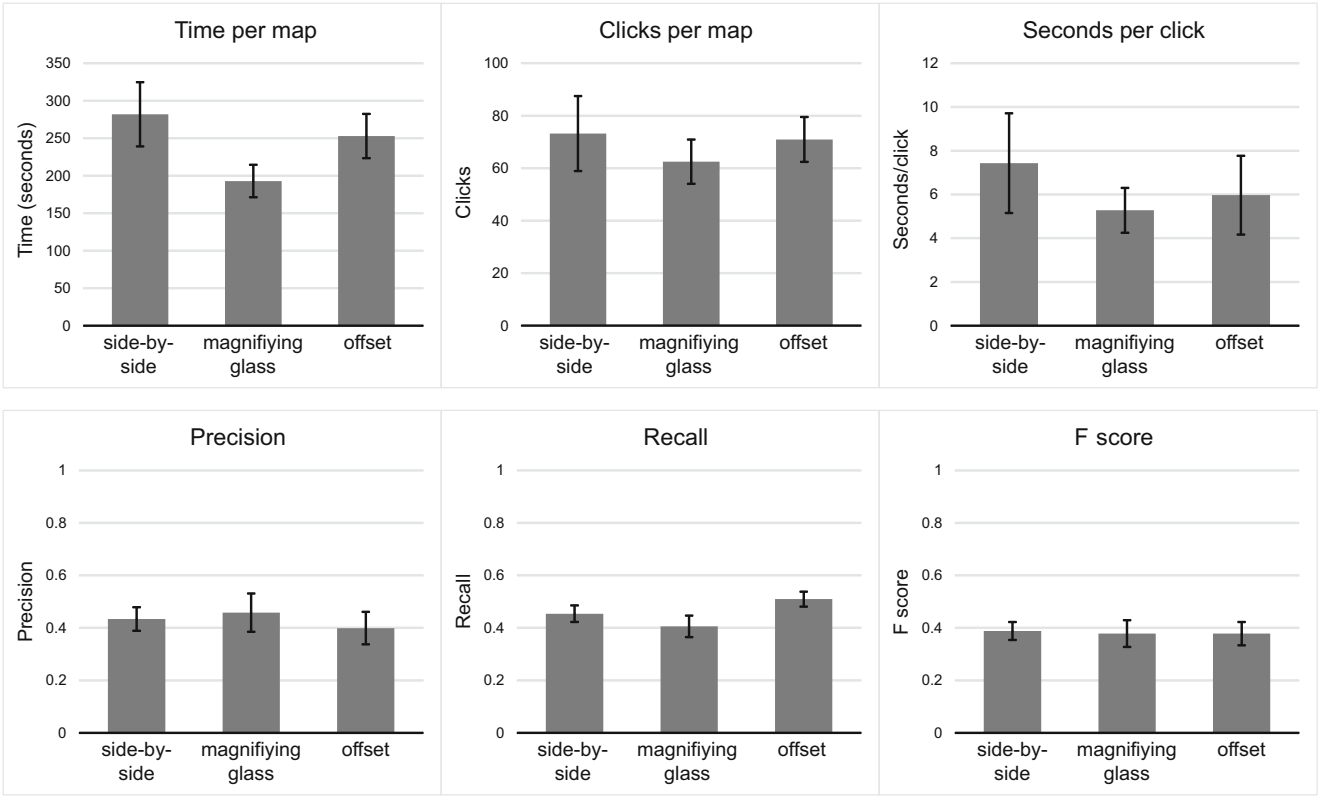

Fig. 3 Comparison of search performance in the three magnification conditions in Experiment 1. Differences are not significant. Error bars show standard error of the mean

Leaving time analysis

First, we looked at whether participants used a giving up time (GUT) strategy to decide when to quit searching a map. The time since last finding a target serves as a measure of the instantaneous rate of target collection; according to marginal value theorem, participants should quit when their rate falls below the average for the whole task, or when their time since last target exceeds the average in the task. The average for the task is computed as the final total of all targets found divided by the total time participants spent in the task. Importantly, this includes the "travel time," the dead time between the final click on one map and the appearance of the next. A plot of the giving up times (time between finding the last target and leaving the map) in our study is shown in Fig. 4. If participants used a giving up time strategy, we would expect these times to be clustered around the average time/target in this task, but this is not the case – giving up times are quite variable, but generally they are shorter than the average time/target. This means that people in this task are not using a simple giving up time strategy: They do not wait until their time since last target exceeds the average time/target.

Next, we looked at whether participants used a potential value strategy to decide when to leave each map. According to this theory, the optimal time to quit a patch is when the expected rate from staying (E1) falls below the expected rate from leaving (E0):

$$E_1 < E_0 \tag{1}$$

The expected rate from leaving a patch is the average rate of target collection (targets/time) in the task environment. The expected rate from staying in a patch varies with the time spent in the patch (t): it's E[X(t)], the expected targets collected at time t, over the total time expended, which is the time in the patch (t) plus the travel time to get to that patch (τ).

$$E_1 = \frac{E\left[X\left(t\right)\right]}{t+\tau} \tag{2}$$

The expected targets at time t depends on the initial number of targets and the nature of the search task. We use the number of clicks on the map as our unit of "time" (t and τ) rather than using a standard measure of time like seconds because it allows us to fit the search task more easily into a probabilistic model. This requires converting the travel time between maps (τ) from seconds to clicks. Since the average loading time between maps was 5 seconds and the average interclick interval during search (averaged over all clicks made in the

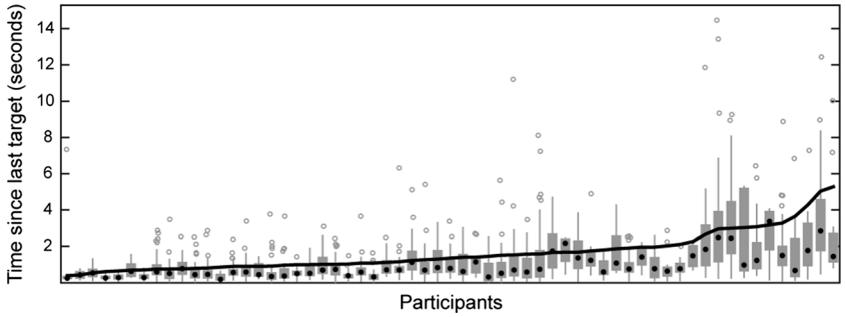

Fig. 4 Boxplots of time between finding the last target and quitting a trial ("Giving Up Time") for each participant in Experiment 1. According to MVT, people should quit when this time reaches their average time between targets, indicated by the solid line. The median leaving times (black dots) for most participants are below the average, which means participants generally quit trials earlier than they should according to MVT

experiment) was 2.5 seconds, we consider travel time τ equivalent to 2 clicks.

In visual search tasks where the target doesn't immediately pop out, it is likely that people process only part of the image at a time. This is often described as a series of deployments of attention over the image. Each deployment selects some subset of items to process and to compare to some target template (Wolfe, 2007). Extending this kind of model to search in natural scenes is difficult, since for purposes of search it's not clear exactly what constitutes an "item" in a scene. It is possible that some groups of objects are processed together as single items, and it is possible that an item (e.g., a face) might be composed of other items (eyes, nose, etc.; Wolfe et al., 2011). Alternatively, one could consider some window around the point of fixation as a surrogate for the item. Thus, as people move their attention around the scene, they would sample a series of windows rather than items per se. However, modeling these from fixations is a difficult problem, since it's not clear how large the window should be; it may depend on the specific search task and how difficult the target is to see against its background.

In the present task, the magnifier provides a useful surrogate for the item or the window around fixation. For the purposes of modeling search in our images, we can use the portion of the scene shown in the magnified view as the equivalent of an item. On each search click, the participant can see a fixed proportion of the map ($1/\omega$) in the zoomed-in view. Thus, we treat that sample as one item from a pool of ω items. (We should note that, in order to simplify the model and reduce the number of free parameters, we have assumed that search windows have a fixed size and do not overlap. This was not entirely true in our experiment – people could change the zoom window size during a trial and sometimes did select overlapping regions – but it is much simpler and not dramatically incorrect to assume a constant window size throughout.) We can compute what cumulative proportion of the map has been searched after each click and how many targets the participant should have found assuming various search strategies. For example, if participants were searching the map exhaustively from left to right and top to bottom, we could model that as random-without-replacement selection from a set of ω items and determine $E[X(t)]$ for some initial number of targets (N) in the display: $E[X(t)] = Nt/\omega$. If a participant had clicked on half the locations ($2t = \omega$), then they would be expected to have found half of the N targets, on average.

However, in our task, people do not search randomly: they prioritize the parts of the image that are most likely to be targets. Returning to Fig. 2, you would not spend clicks on the golf course, and you would be unlikely to spend many on the areas that appear to contain only residential housing. Since participants guide their attention and their clicks to areas deemed most likely to contain gas stations and search those areas first, their expected rate of target collection starts out higher and falls off much faster than it would if search were simply random.

The exact shape of that function depends on how efficient the search task is. For example, in a very efficient "pop-out" search task, such as collecting red targets among green distractors, people could collect all of the targets immediately: Their expected rate would be one target per click until the targets were exhausted, at which point their expected rate would fall to zero. The gas station search task falls somewhere in between this very efficient, perfectly guided search and a random search.

It would be difficult to say, in theory, how efficient gas station search should be, but we can estimate it directly from our search data. We assume that people in our task have a way of deciding which sections of the image are most likely to contain targets, and that they generally search regions in order from most to least likely, with some noise. We can estimate search efficiency by looking at how the number of targets found relates empirically to the proportion of the image searched. We divide each image into a 25 × 25 grid (approximately the size of the average zoom window), giving us a surrogate set size of 625 items/regions. Next, we make a histogram of all participants' search clicks on that image, shown graphically in Fig. 5b, with hotter colors indicating more clicks in that element of the 25 × 25 grid. It can be seen that, in these images, search clicks cluster

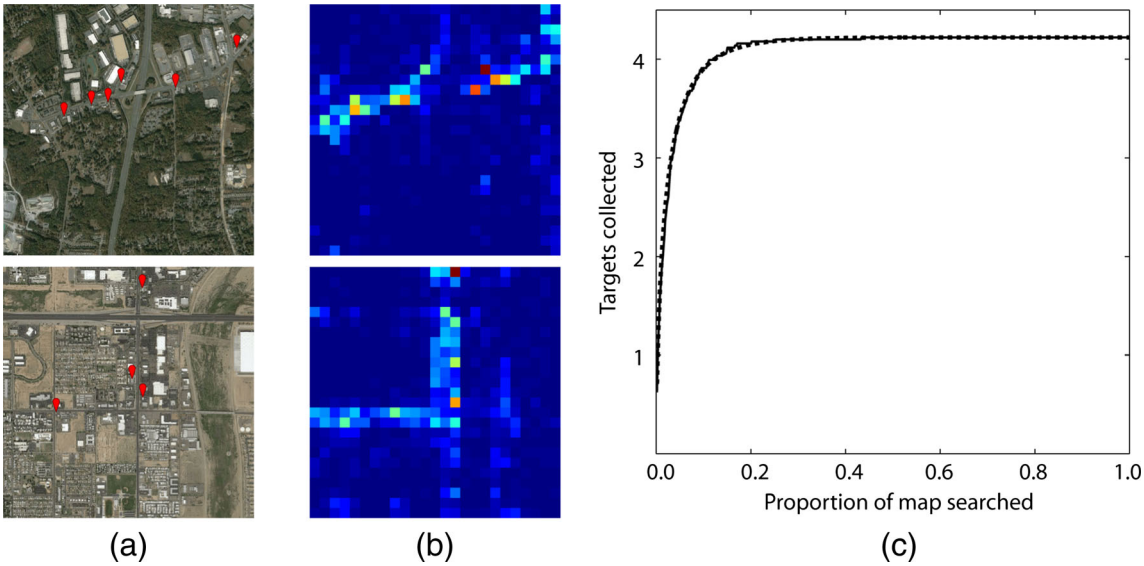

Fig. 5 Estimating the rate of target collection in the map search task. (**a**) Two example maps, red markers indicate gas stations. (**b**) Heatmaps showing the distribution of search clicks in these maps, averaged over all subjects. Pixels colored red are areas that were searched most frequently; dark blue areas were search least frequently. (**c**) Cumulative targets per area of the map searched, assuming search from most-frequent to least-frequent locations, averaged over maps with at least one target. The solid line is the empirical curve from the heatmaps in (**b**); the dotted line is the fitted curve used in modeling. Map imagery © Google, Digital Globe and Orbis, Inc. (Color figure online)

along main roads, lined with buildings. There is very little search for gas stations in empty fields or near stretches of limited-access highway. Now we can take the top N% of grid elements and ask how many targets fall in this area. If we sweep from 0% to 100% of the grid, the resulting curve is shown in Fig. 5c. Figure 5c is the average of these curves across all images. The *y*-axis of this curve runs from zero to about four gas stations because the average number of gas stations in our set of images was about four. Naturally, there is some variation in search speed across maps – gas stations can be found quickly in some, while others require much greater scrutiny – but this curve gives the average rate of target collection for this task. From this curve, we can see that, even though this is a difficult, slow task, participants were very efficient in terms of the proportion of the image scrutinized. Once they had examined about 9% of the map area, on average, they would have found 90% of the available gas station targets. This illustrates the very "guided" nature of this search task. Participants were able to use their understanding of scene context and their knowledge of the rough visual features of the targets to guide their search to the buildings most likely to be gas stations.

Knowing the expected rate of target collection allows a forager to choose the optimal time to quit a guided search task. It also allows a forager to refine their expectations about the number of targets in a patch, since the expected rate also depends on the number of targets in the patch. By comparing the actual number of targets collected to the expected collection rates for different numbers of targets initially in the patch, a forager can determine what target count is most probable.

How the expected rate of target collection varies with the number of targets depends on how strongly guided the search is. If search is completely random, the odds of finding a target on each click is directly proportional to the number of targets available: If the odds of finding a target on the first search click is p in an image with one target, the odds of finding a target on the first click in an image with 10 targets is 10p. However, this is not quite correct for guided search: If the odds of finding a target on the first search click is 95% in an image with one target, the odds of finding a target on the first click in an image with 10 targets is higher, but it's not 950%. Understanding how the search expectations vary with the number of targets in the image requires representing guided search as a signal detection problem.

Let us assume that, when deciding where to click next on each map, participants must make a two-alternative, forced-choice decision about whether a region will contain a target or not. This decision is based on a "targetness" signal that is computed from local visual features that resemble the target and spatial location information that suggests where a target is most likely to be present. For example, if the search task is to find blue cars in a parking lot scene, then image locations that have blue colors and are in the bottom half of the image (not the sky) are the most likely target locations. This targetness signal probably wouldn't be perfect for most search tasks. If the signal is thresholded, then search actions (fixations or clicks) guided by this targetness signal can be classified into hits or false alarms according to whether or not they land on a target. For example, in the car search task, the viewer's first

two fixations might fall on a blue mailbox and a blue car: The former could be considered a false alarm and the latter a hit. In fact, we can think of search efficiency as reflecting how well target regions and distractor regions (meaning, regions which do not contain a target) are separated by the targetness signal (Wolfe, 2007). Again, for simplicity, we assume that regions do not overlap and either are or are not targets.

We assume that the target regions and distractor regions come from two overlapping normal distributions on this targetness scale. In an easy guided search task, such as finding a red dot among green, these distributions would be very well separated. In a very difficult search task, such as identifying cancer in a mammogram, the target and distractor distributions may overlap quite a bit – this reflects the difficulty in determining at a glance whether a given region in the image contains a target and guiding attention to the most likely regions. Using the curve in Fig. 5c, we can determine the amount of overlap between targets and distractors in our task, and how efficiently people were able to search these scenes. We scale this average curve so that the maximum number of targets is one so we can treat it as an ROC curve. An ROC curve plots the percentage of hits in a two-alternative forced-choice task against the percentage of false alarms. In this case, we treat *every* search click that does not reveal a target as a false alarm, since we assume that people search these images selectively and only click on regions that have a reasonably high probability of containing a target. (This broader definition of "false alarm" includes the traditional false alarms – patches that the observer clicked and then incorrectly marked as targets – but these are only a small percentage of the unsuccessful search clicks. In most cases, the observer could identify a distractor patch as a distractor without marking it.) We then fit a binormal function (Tourassi, 2012), which generates a theoretical ROC curve for a two-alternative forced choice between two normal distributions. In other words, we treat Fig. 5c as the ROC curve for human participants who classify regions of satellite images as "target" (contains gas station) versus "distractor" (contains no gas stations). We use a standard ROC fitting technique to derive the parameters of the target and distractor distributions used for this task. The binormal fit gives parameters $\alpha = 3.41$ and $\beta = 1.57$, which relate to the means and standard deviations of the target and distractor distributions as follows:

$$\alpha = \frac{\mu_{distractor} - \mu_{target}}{\sigma_{distractor}};\ \beta = \frac{\sigma_{target}}{\sigma_{distractor}} \tag{3}$$

The parameter α is equivalent to d' when the standard deviations of the two groups are the same, and the parameter β is the ratio of standard deviations for the two groups. By setting the mean of one group to zero and the standard deviation of the other group to 1, α and β give the mean and standard deviation of the other group. In this case, we set the target distribution to have $\mu = 0$ and $\sigma = \beta$ and the distractor distribution to have $\mu = \alpha$ and $\sigma = 1$. Note that this means the target distribution has the lower mean (this is simpler for later computations), so the signal used to distinguish targets from distractors can be thought of as a "distractor-ness" signal: image regions with lower values are more likely to be targets, and a guided search for targets would proceed from the lowest ranked regions to the highest (left to right in Fig. 6).

Given these parameters, we can predict the search curve for a display with any number of targets. We scale the target and distractor distributions according to the number of targets N: The target distribution has area N, and the distractor distribution has area (ω - N). We define the number of "items" or samples in the display (ω) by the size of the average zoom window. In our experiment, the average zoom level was 5.29×, which was about 1/606 the area of the map (note that since 606 is not a square number, we rounded to the nearest square number, 625, when building the grid in the previous step). We can model the guided search process as sampling image regions from these overlapping distributions, in order, from most target-like to least target-like (left-to-right in Fig. 6). The expected targets at a given sample t is the cumulative area under the target distribution function at the point where the total cumulative area under both distributions is t. We denote this point as γ_t. This cumulative area represents the likelihood of collecting targets rather than distractors: The more well-separated the two distributions, the more targets should be collected in the earliest stages of the search. The formula for this is as follows (Φ represents the cumulative distribution function of the standard normal distribution):

$$E[X(t)] = \mathrm{N}\Phi\left(\frac{\gamma_t}{\beta}\right) \tag{4}$$

The value γ_t can be calculated numerically from:

$$t = N\Phi\left(\frac{\gamma_t}{\beta}\right) + (\omega - N)\Phi(\gamma_t - \alpha) \tag{5}$$

The expected targets, and therefore the expected rate, depends on the initial number of targets in the patch (N): Regardless of search efficiency, if there are more targets available, more of them should be found at each point in the search. In our task, searchers don't know how many targets there will be in each patch, so we treat N as a probability distribution over possible numbers of targets:

$$E_1 = \sum_{n=0}^{n_{max}} p(N = n)\frac{E\left[X(t)\middle|N = n\right]}{t + \tau} \tag{6}$$

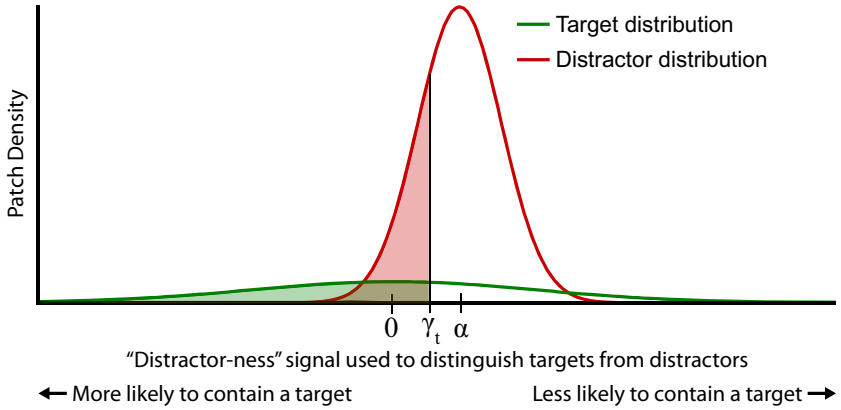

Fig. 6 Signal detection model for guided search. We assume that people have a priority map of the image in which regions are ranked from most likely to contain targets to least likely to contain targets (*x*-axis). Search involves sampling these regions in order from most to least likely. The distributions represent the likelihood that a region is actually a target or distractor. The area under the target curve is the number of image patches that contain targets, and the area under the distractor curve is the number of patches in the image without targets. The expected number of targets after searching a given number of patches (t) can be determined by finding the point on the *x*-axis where the sum of the cumulative distributions (shaded area) equals t (we call this point γ_t) and taking the cumulative distribution of the target curve up to that point (green shaded area). (Color figure online)

We use the responses from the "How many gas stations" guessing task as the prior p(N). We make a histogram of the responses, smooth it with a Gaussian with standard deviation 1.5, truncate it at n_{max} and normalize it so it sums to 1. We arbitrarily chose a value of 20 for n_{max} (17 was the highest target count guess for any map).

Equation 6 could be used to compute the expected rate in the patch at any sample t, assuming that the prior p(N) doesn't change as the forager searches the patch. This might be true for certain tasks where the forager knows or can accurately guess the number of targets in a patch before searching it, but it probably isn't true in this task because participants start with only a rough idea of the number of gas stations in each map. A smart searcher should update that estimate on the basis of their experience searching the map. Therefore, we use a Bayesian updating step to determine the probability on the number of targets at sample t, $p(N^{(t)} = n)$, based on the observed search result *obs*, a binary variable that represents whether or not the searched location is a target. The posterior probability on N (meaning: the updated beliefs about the true initial number of targets) comes from Bayes' rule: It's the probability on N from the previous sample times the likelihood of the search result for a given number of targets:

$$p\left(N^{(t)} = n \middle| obs\right) \propto p\left(obs \middle| N^{(t-1)} = n\right) p\left(N^{(t-1)} = n\right) \quad (7)$$

The likelihood of the search result can be determined from the target and distractor distributions described previously. Suppose the summed distribution is divided into ω samples, each with area equal to 1. To determine the odds of finding a target on a given sample t, we look at the area under the distribution between γ_{t-1} and γ_t (the cumulative distribution between sample t and sample t-1, computed from Eq. 5). The probability of finding a target on sample t is the area under the target distribution within this window, and the probability of not finding a target is 1 minus this value.

$$p\left(target \middle| N^{(t-1)} = n\right) = N\left(\Phi\left(\frac{\gamma_t}{\beta}\right) - \Phi\left(\frac{\gamma_{t-1}}{\beta}\right)\right) \quad (8)$$

$$p\left(no\ target \middle| N^{(t-1)} = n\right) = 1 - p\left(target \middle| N^{(t-1)} = n\right) \quad (9)$$

Finally, the threshold leaving time E_0 is estimated from our data: It's the total targets over total clicks, averaged for each participant. Since we measure time in the map by search clicks, we also need to specify the travel time between maps in clicks. As noted above, we use a travel time of two clicks.

To summarize, the potential value model assumes that people make their decision to quit based on how many targets they believe are in an image and how quickly they should be able to find them. While searching, they update these beliefs in a Bayesian fashion – so, for example, if targets are harder to find than expected, people may conclude that the image contains fewer targets than they initially thought. People use their beliefs about the number of targets available and the expected search efficiency to determine their expected rate of target collection in that image, and when that expected rate falls below the average rate for the task, they quit searching and move on to the next image.

In our model, the prior on the number of targets in each image and the expected search efficiency are set for each image. On each trial, we use the image priors and an individual participant's search history – how many targets were found and the time taken to find them – to estimate the participant's expected rate when quitting that trial. That search history is the only individual participant data used in the model; all other

parameters are constants, or constant for a particular image (e.g., the prior on the number of targets in that image). We compare the predictions of this full model to the predictions of MVT, which says that participants quit when their instantaneous rate, based on the time since collecting the last target, falls below the average rate. For each participant on each trial, we find the instantaneous rate upon quitting the trial by taking one over the time elapsed since finding the last target (1/GUT). Like the potential value model, MVT uses the individual participant's search history in a single trial to compute leaving times, but unlike the potential value model, it only considers the last target collected. Finally, we compare the full potential value model to a "prior-only" version, which uses the prior on the number of targets in each image and the expected search efficiency but does not update these beliefs based on the participant's search history. On each trial, this model estimates the participant's expected rate based on image priors alone without using any information about the participant's search history on that trial.

The comparison of the three models is shown in Fig. 7. Each panel in Fig. 7 plots the median instantaneous or expected rate when leaving a map against the average target collection rate for each participant in Experiment 1. The instantaneous rate, measured as 1/GUT in Fig. 7a, is not a bad predictor of leaving time in the sense that that the median rate when leaving the map is correlated with the average rate ($r = 0.67$, $p < .01$). However, the prediction, while correlated, is not accurate. The values are quite variable, and instantaneous rates are virtually always higher than the average rates. Thus, this model fails because people are leaving before the instantaneous rate drops to the average rate, against the prediction of a standard marginal value account.

Figure 7b and c show results for two models based on the potential value theorem. According to the potential value theorem, people should leave a map when their *expected* rate of target collection falls to the average rate. Figure 7b shows a "prior-only" model that only uses the prior on the number of targets (from the guessing task) to predict the expected rate in the map (Eq. 6). Figure 7c shows the "full model" that uses participants' search results to update beliefs about the number of targets after each click on the map (Eqs. 7–9). The expected rate, as computed by the "prior only" model is less well correlated with the average rate than the marginal value model ($r = 0.56$, $p < .01$), which suggests this model is a poorer predictor of when participants will leave an image. The full model, however, seems to predict leaving times rather well: The median rate is well correlated with the participant's average rate ($r = 0.82$, $p < .01$). This means that, in general, participants leave a map when their expected rate (based on their prior beliefs about the map and their search experience) falls to their average rate, as predicted by potential value theorem.

Comparisons of the median leaving rate versus average rate from the three models are shown in Table 1. Mean difference is computed by taking the difference between the model's leaving rate and average rate (predicted – observed) for each participant. The mean and highest density intervals for this distribution are computed using Bayesian estimation (BEST) with the methods and default parameters described by

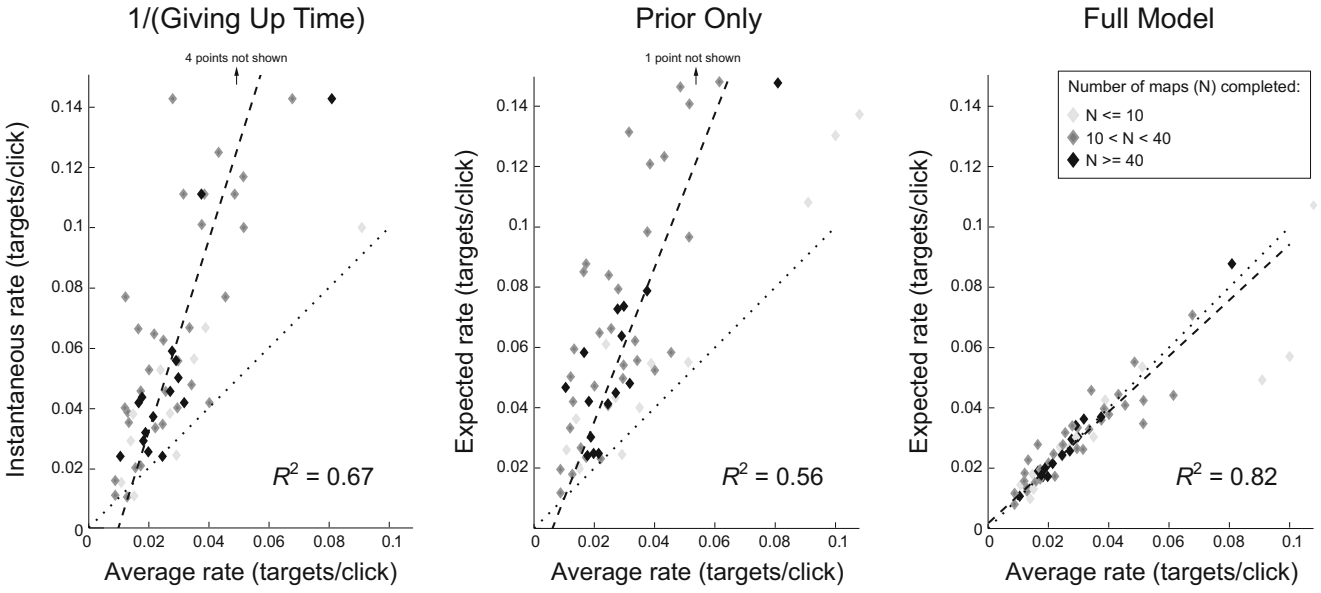

Fig. 7 Plots of participants' median rate when quitting trials vs. their average rates in Experiment 1, for various models. Each diamond represents a participant, with shade coding the number of maps that participant completed (out of 50). The dotted line indicates the identity (1:1) line: if participants left the patch when their expected rate was exactly the average rate, then all points would lie along this line. Dashed lines are the best linear fit to the data; R^2 for each fit is given on the graph

Table 1 Model comparison

Model	Mean difference (95% HDI)	AIC
Instantaneous rate (1/GUT)	0.0242 (0.0163, 0.033)	-358.84
Expected rate from prior only	0.0313 (0.0233, 0.0392)	-341.92
Expected rate from full model	0.00091 (0.0001, 0.0019)	-397.21

Note. HDI = highest density interval; AIC = Akaike information criterion

Kruschke (2013). All of the models have mean differences significantly above zero (the 95% highest density interval of the mean does not include zero), but the full model has the lowest mean difference. Planned comparisons using one-group BEST show that the mean difference for GUT is significantly higher than the mean difference for the full model (estimated mean difference = 0.0233, 95% highest density interval, HDI, = [0.0158, 0.0315]) and the mean difference for the prior-only model is significantly higher than the mean difference for the full model (estimated mean difference = 0.0332, 95% HDI = [0.0239, 0.0404]). The mean difference for the prior-only model is higher than the mean difference for GUT (estimated mean difference = 0.0047, 95% HDI = [0.000523 0.00884]).

We also used linear regression to predict the average leaving times from the median instantaneous or expected rates. This analysis does not assume that leaving rates should exactly match average rates, but it does assume they should be consistently, linearly related (e.g., participants might leave when their rate is half the average). Akaike information criterion (AIC) for each model is given in Table 1. This is a measure of model fit; lower values indicate a more probable model. The relative likelihood can be used to determine the significance of a difference between two models' AIC: Relative likelihood indicates the probability that the model with lower AIC is actually better than a model with higher AIC. Relative likelihood is computed as $\exp((AIC_2 - AIC_1)/2)$, where AIC_1 is the lower AIC value. The GUT model is a significantly better fit to the data than the prior-only model: The relative likelihood of GUT compared to the prior-only model is 4,722. However, the full model is a significantly better fit than either of these models: The relative likelihood of the full model compared to GUT is 2.1E+8 and the relative likelihood of the full model compared to the prior-only model is 1.0E+12.

Experiment 2

In Experiment 1, participants were not required to complete all of the maps and could quit the task whenever they wanted. We wished to determine if this freedom would produce different behavior from a version of the experiment where one was required to complete a fixed number of maps. Thus, Experiment 2 was a replication of Experiment 1 in which we equalized the number of trials per subject and number of subjects per condition. Each participant viewed exactly 24 maps, and all participants viewed the same 24 maps. In order to test the generality of our model, the parameters, derived in Experiment 1, were used to model performance in Experiment 2.

Method

Participants

Thirty-six people participated in Experiment 2 on Amazon Mechanical Turk; none had participated in Experiment 1. The participant requirements and consent procedure were identical to Experiment 1. In Experiment 2, participants received a base payment of $6.00 for finding at least 30 gas stations and a bonus of $0.20 for each gas station after the first 30.

Stimuli

Twenty-four of the 50 maps from Experiment 1 were used in Experiment 2. The maps were selected to give a more uniform distribution of target counts: 1, 2, 3, 4, 5, 6, 7, or 9 targets (three maps of each type).

Design and procedure

The viewing conditions, map interface, and search task procedure were identical to Experiment 1, except that we did not ask participants to give confidence ratings after marking potential targets. Participants were randomly assigned to one of the three viewing conditions (eight participants per condition), and all participants were required to search all 24 maps.

Results and discussion

We dropped 13 trials over 60 minutes in length and 34 trials with no clicks recorded, leaving 817 trials. Most (83%) of these dropped trials were from participants in the magnifying glass condition, but there were no outlier participants with an unusually high number of dropped trials. As in Experiment 1, we compared the three magnifier conditions in a by-subjects analysis, shown in Fig. 8. There was no significant difference in the total time per map, ($F(2,33) = 1.79$, $p = 0.18$), number of search clicks per map, $F(2, 33) = 0.53$, $p = .59$, rate of clicks, $F(2, 33) = 1.44$, $p = .25$, average recall, $F(2, 33) = 1.65$, $p = .21$, average precision, $F(2, 33) = 1.50$, $p = .23$, or F score, $F(2, 33) = 1.26$, $p = .30$. This replicates our nonsignificant findings from Experiment 1: The different magnifying interfaces do not seem to affect search in this task.

To investigate leaving times in this task, we used the model described in Experiment 1. We kept all of the model parameter

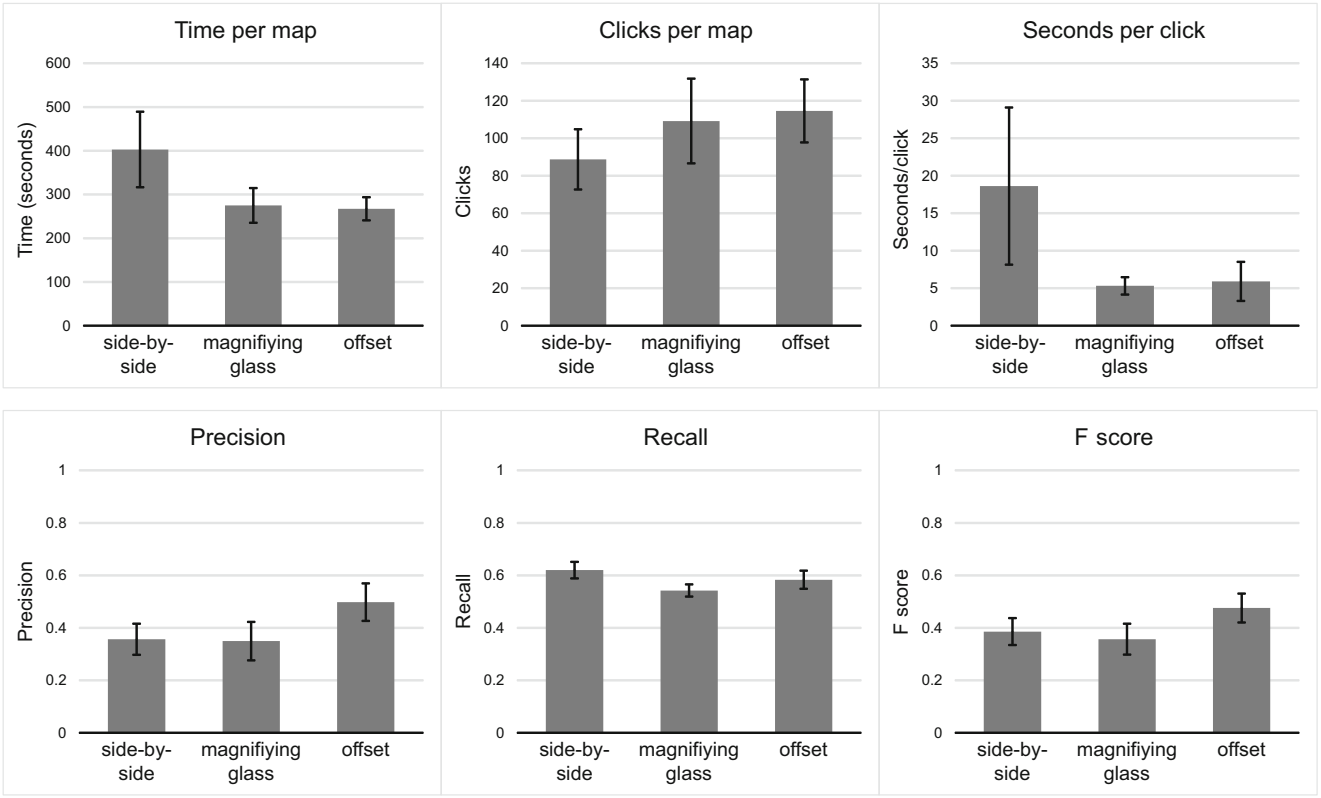

Fig. 8 Comparison of the three magnification conditions in Experiment 2. Differences are not significant. (Although the side-by-side condition seems to have a much slower click rate than the other conditions, this is due to a single outlier participant.) Error bars show standard error of the mean

values (ω, α, β, τ) that we had computed from the Experiment 1 data but used them to predict the search behavior observed in Experiment 2. As in Fig. 7, Fig. 9 shows the three different calculations of the median rate when quitting a map plotted against the average rate for each participant in Experiment 2. As in Experiment 1, instantaneous rate (1/GUT) was least well correlated with average rate (r = 0.81, $p < .01$). Expected rate from priors only was better correlated (r = 0.85, $p < .01$), but, like the 1/GUT measure, the priors only measure predicts that observers will leave a map sooner than is the case. The expected rate based on priors and search experience was very well correlated with the average

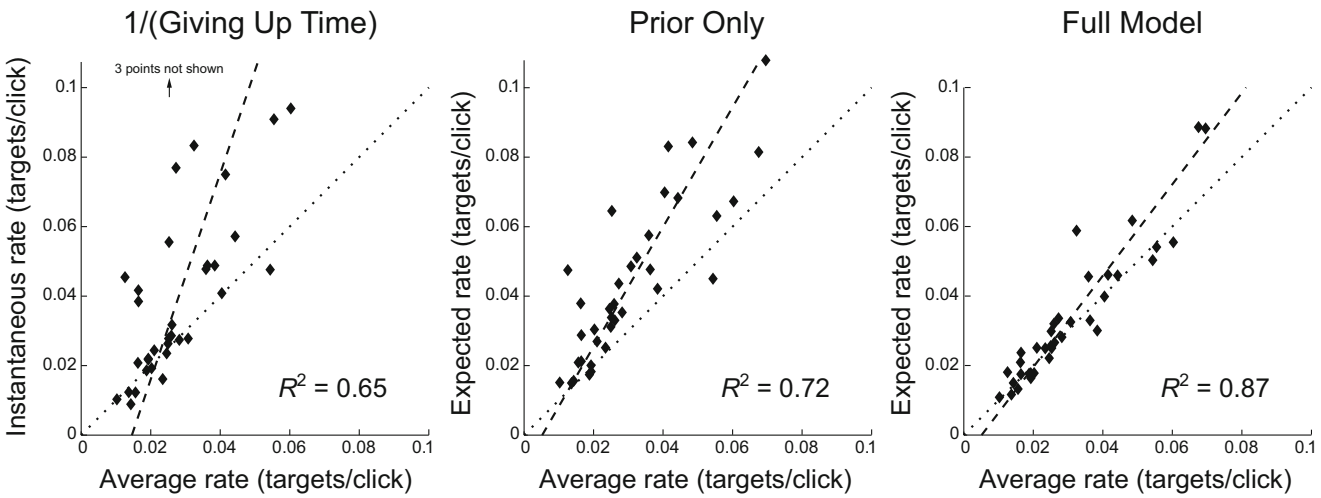

Fig. 9 Plots of participants' median rate when quitting trials vs. their average rates in Experiment 2, for various models. The dotted line indicates the identity (1:1) line: If participants left the patch when their expected rate was exactly the average rate, then all dots would lie along this line. Dashed lines are the best linear fit to the data; R^2 for each fit is given on the graph

rate (r = 0.93, p < .01). Mean difference and AIC for each model are given in Table 2. Planned comparisons using one-group BEST show that the mean difference for GUT is significantly higher than the mean difference for the prior-only model (estimated mean difference = 0.057, 95% HDI = [0.0287, 0.0845]) and the mean difference for the prior-only model is significantly higher than the mean difference for the full model (estimated mean difference = 0.021, 95% HDI = [0.0152, 0.0274]). Comparing AIC values shows that the relative likelihood of the prior-only model compared to GUT is 8.4, so these models may not be significantly different (Burnham, Anderson, & Huyvaert, 2011). The relative likelihood of the full model compared to GUT is 6.6E+10 and the relative likelihood of the full model compared to the prior-only model is 7.9E+9, so the full model does seem be a significantly better fit than the other two models.

As in Experiment 1, if we assume that people compute expected rate using both prior beliefs about the map and their experience when searching the map, then their decision on when to quit each trial in this task appears to follow an optimal foraging strategy: They quit when their expected rate of target collection on a map falls to the average rate.

General discussion

We investigated foraging behavior in a task where people used a magnifier interface to search for gas stations in large satellite images and found that quitting times were well predicted by a potential value version of the optimal foraging model. This model predicts that a rate-maximizing forager should leave a patch when the expected rate of target collection in the patch falls below the average rate of target collection for the environment. The expected rate of target collection can't be directly observed by the forager. Foragers must estimate this rate based on their beliefs about the likely number of targets in the patch and the rate at which they should be able to find them. Modeling these beliefs in simple, random search tasks is straightforward, so most previous work has focused on this type of search. Here, we show that the same models can be extended to a highly guided search task where people do not search randomly but prioritize regions or objects most likely to be targets.

Table 2 Model comparison

Model	Mean difference (95% HDI)	AIC
Instantaneous rate (1/GUT)	0.0861 (0.0563, 0.117)	-222.42
Expected rate from prior only	0.0300 (0.0221, 0.0377)	-226.67
Expected rate from full model	0.00826 (0.00566, 0.0107)	-272.26

Note. HDI = highest density interval; AIC = Akaike information criterion

We describe two methods for estimating the expected rate in a patch. In both cases, we assume that people have some prior expectations about how many targets will be present in an image. In the prior-only model, we assume that people only use this prior and average search curve, which gives them an estimate of how quickly they should be able to find targets in the image. Across our experiments, this approach performs about as well as the marginal value approach, which assumes that leaving times are based on the time since last finding a target. Like the marginal value account, the leaving times estimated according to the prior-only account are quite variable and generally come earlier than would be predicted by optimal foraging. Our full model assumes that people also update their beliefs about the number of targets in a display as they search. Finding targets convinces them that a display is more target-rich than expected, and not finding targets convinces them that the display is poor. This model predicts people's leaving times quite well, which suggests that people are combining all three sources of information (prior expectations, average search efficiency, and their own current search results) to decide when to quit searching the images in this task.

One concern with this task, and many other foraging experiments, is that we can't be certain what the participants were trying to maximize, so it's not clear what the "optimal" strategy would be. A participant who is trying to find all of the targets, for example, would have a different threshold for quitting a map than one who is trying to maximize the rate of target collection. These experiments were run on Amazon's Mechanical Turk, a site where people do short computer-based tasks for money. It seems likely that the average worker on the site is trying to maximize his or her hourly wage, which in our task would mean maximizing the rate of target collection. That said, there are many reasons why a worker might do something nonoptimal in this kind of task. Participants may have their own subjective cost functions that includes factors other than hourly wage: For example, they might keep searching a map longer than "optimal" because they really dislike missing targets or because they enjoy doing the task for its own sake. (About 40% of our participants left us feedback about the task and, perhaps surprisingly, the majority described it as "fun" or "enjoyable.") It's also important to note that, since the number of maps in our task was limited, workers weren't only making a choice between continuing to search the current map or moving on to the next map. To some extent, they were trading off time on the current map for time they could spend on another Mechanical Turk task. So the true threshold for a wage-optimizing worker is actually the average rate of pay on the Mechanical Turk site as a whole (which is difficult to determine, but probably similar to the average rate of pay in our task).

Although this study focuses on a task with multiple targets, the potential value approach is also applicable to standard

search tasks where there is only one target that is either present or absent. In the standard case, where targets are randomly present on half of trials and there are no image priors to guide search, the potential value model prediction is similar to other models of quitting time in search: People should give up the search once their belief that a target is present (and the expected rate is greater than zero) falls below a threshold. This giving up time would depend on the expected search efficiency, so it would be longer for more difficult searches. In cases where there are different priors on target presence for different images (e.g., search for real objects in natural scenes), the potential value model would generally predict longer search in images with higher priors.

One place where the potential value model may be useful is in modeling prevalence effects in search. Previous work has show that when search targets are uncommon, they are more likely to be missed (Wolfe & Horowitz, 2007). In a potential value model, that means that the prior on target presence is low, and people may be able to decide that an image is target-absent with less search evidence than they would need to make the same decision when targets are more common. However, the current study did not include any extreme manipulations of target prevalence, so further research would be needed to determine whether this is the case.

In addition to looking at foraging behavior in this task, we also examined the effect of different magnification interfaces. We found that the task interface had no significant effect on search performance, contrary to a previous study (Zhao et al., 2009) which showed that an embedded magnifier is easier to use than one which shows a zoomed-in view off to the side. However, the search task in their study was very different: They used a word search whereas we looked at search in natural images. Our scenes had a coherent structure with many landmark features such as roads and rivers which probably helped searchers navigate through the images and keep track of what areas they had already searched. We also used a large, salient footprint in the overview map to help users keep track of the zoom window's location in the side-by-side condition, a feature which the Zhao et al. (2009) interface didn't include. This may have made the side-by-side magnifier easier to use so that it was not significantly worse than the embedded magnifiers.

The requirement to use the magnifier to confirm the presence of a target provides a novel way to look at guidance in a complex, extended search task. The method works because there is enough information in the scene to guide attention but not enough to identify the target. This method could be used studies of other complex search stimuli, giving us new insight into the way that our knowledge interacts with a stimulus in order to make search reasonably efficient.

Acknowledgments This work was funded by a National Geospatial Agency grant to J. M. W.

References

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision, 9*(12), 1–18.

Bergen, J. R., & Julesz, B. (1983). Rapid discrimination of visual patterns. *IEEE Transactions on Systems, Man and Cybernetics, 13,* 857–863.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology, 65,* 23–35.

Cain, M. S., Vul, E., Clark, K., & Mitroff, S. R. (2012). A Bayesian optimal foraging model of human visual search. *Psychological Science, 23*(9), 1047–1054.

Chan, L. K. H., & Hayward, W. G. (2013). Visual search. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(4), 415–429.

Charnov, E. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology, 9*(2), 129–136.

Chun, M. M., & Wolfe, J. M. (1996). Just say no: How are visual searches terminated when there is no target present? *Cognitive Psychology, 30*(1), 39–78.

Cousineau, D., & Shiffrin, R. M. (2004). Termination of a visual search with large display size effects. *Spatial Vision, 17*(4), 327–352.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96*(3), 433–458.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision, 11*(5), 1–36.

Egeth, H. E., Jonides, J., & Wall, S. (1972). Parallel processing of multielement displays. *Cognitive Psychology, 3,* 674–698.

Egeth, H. E., Virzi, R. A., & Garbart, H. (1984). Searching for conjuctively defined targets. *Journal of Experimental Psychology: Human Perception & Performance, 10*(1), 32–39.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition, 17,* 945–978.

Green, R. F. (1980). Bayesian birds: A simple example of Oaten's stochastic model of optimal foraging. *Theoretical Population Biology, 18,* 244–256.

Green, R. F. (1984). Stopping rules for optimal foragers. *The American Naturalist, 123*(1), 30–43.

Hills, T. T., Todd, P. M., & Jones, M. N. (2015). Foraging in semantic fields: How we search through memory. *Topics in Cognitive Science, 7*(3), 513–534.

Krebs, J. R., Ryan, J. C., & Charnov, E. L. (1974). Hunting by expectation or optimal foraging? A study of patch use by chickadees. *Animal Behaviour, 22,* 953–964.

Kruschke, J. K. (2013). Bayesian Estimation Supersedes the t Test. *Journal of Experimental Psychology: General, 142*(2), 573–603.

Kundel, H. L., & La Follette, P. S., Jr. (1972). Visual search patterns and experience with radiological images. *Radiology, 103*(3), 523–528.

Kwak, H. W., Dagenbach, D., & Egeth, H. (1991). Further evidence for a time-independent shift of the focus of attention. *Perception & Psychophysics, 49,* 473–480.

McNamara, J. (1982). Optimal patch use in a stochastic environment. *Theoretical Population Biology, 21,* 269–288.

McNamara, J., Green, R. F., & Olsson, O. (2006). Bayes' theorem and its applications in animal behaviour. *Oikos, 112,* 243–251.

McNamara, J., & Houston, A. (1985). A simple model of information use in the exploitation of patchily distributed food. *Animal Behaviour, 33,* 553–560.

Moran, R., Zehetleitner, M. H., Müller, H., & Usher, M. (2013). Competitive guided search: Meeting the challenge of benchmark RT distributions. *Journal of Vision, 13*(8), 1–31.

Neider, M. B., & Zelinsky, G. J. (2008). Exploring set-size effects in scenes: Identifying the objects of search. *Visual Cognition, 16*(1), 1–10.

Oaten, A. (1977). Optimal foraging in patches: A case for stochasticity. *Theoretical Population Biology, 12,* 263–285.

Palmer, J. (1995). Attention in visual search: Distinguishing four causes of a set-size effect. *Current Directions in Psychological Science, 4*(4), 118–123.

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research, 40*(10–12), 1227–1268.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review, 106*(4), 643–675.

Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General, 109*(20), 160–174.

Stephens, D. W., Brown, J. S., & Ydenberg, R. C. (2007). *Foraging: Behavior and Ecology.* Chicago, IL: University of Chicago Press.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory.* Princeton, NJ: Princeton University Press.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*(4), 766–786.

Tourassi, G. (2012). ROC analysis: Basic concepts and practical applications. In S. Ehsan & E. Krupinski (Eds.), *The handbook of medical image perception and techniques* (pp. 187–203). New York, NY: Cambridge University Press.

Townsend, J. T. (1971). A note on the identifiability of parallel and serial processes. *Perception & Psychophysics, 10*(3), 161–163.

Võ, M. & Wolfe, J. M. (2015). The role of memory for visual search in scenes. *Annals of the New York Academy of Sciences, 1339*, 72–81.

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 1,000,000 Trials Tell Us About Visual Search? *Psychological Science, 9*(1), 33–39.

Wolfe, J. M. (2003). Moving towards solutions to some enduring controversies in visual search. *Trends in Cognitive Sciences, 7*(2), 70–76.

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 99–119). New York, NY: Oxford.

Wolfe, J. M. (2012). When do I quit? The search termination problem in visual search. *Nebraska Symposium on Motivation, 59,* 183–208.

Wolfe, J. M. (2013). When is it time to move to the next raspberry bush? Foraging rules in human visual search. *Journal of Vision, 13*(3), 1–17.

Wolfe, J. M. (2014). Approaches to Visual Search: Feature Integration Theory and Guided Search. In A. C. Nobre & S. Kastner (Eds.), *Oxford Handbook of Attention* (pp. 11-55). New York: Oxford U Press.

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R. E., & Kuzmova, Y. I. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception, & Psychophysics, 73,* 1650–1671.

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), 419–433.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5,* 1–7.

Wolfe, J. M., & Horowitz, T. S. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *The Journal of Experimental Psychology: General, 136*(4), 623–638.

Wolfe, J. M., Horowitz, T. S., & Palmer, E. M. (2010). Reaction time distributions constrain models of visual search. *Vision Research, 50*(14), 1304–1311.

Wooding, D. S., Roberts, G. M., & Phillips-Hughes, J. (1999). Development of the eye-movement response in the trainee radiologist: Image perception and performance. *Proceedings—SPIE Medical Imaging, 3663,* 136–145.

Zhao, Z., Rau, P.-L. P., Zhang, T., & Salvendy, G. (2009). Visual search-based design and evaluation of screen magnifiers for older and visually impaired users. *International Journal of Human-Computer Studies, 67,* 663–675.