European Journal of Cognitive Psychology, in press

Eye movements during scene inspection: A test of the saliency map hypothesis

Geoffrey Underwood, Tom Foulsham, Editha van Loon,
Louise Humphreys and Jackie Bloyce

University of Nottingham, UK

What attracts attention when we inspect a scene? Two experiments recorded eye movements while viewers inspected pictures of natural office scenes in which two objects of interest were placed. One object had low contour density and uniform colouring (a piece of fruit), relative to another that was visually complex (for example, coffee mugs and commercial packages). In each picture the visually complex object had the highest visual saliency according to the Itti and Koch algorithm. Two experiments modified the task while the pictures were inspected, to determine whether visual saliency is invariably dominant in determining the pattern of fixations, or whether the purpose of inspection can provide a cognitive override that renders saliency secondary. In the first experiment viewers inspected the scene in preparation for a memory task, and the more complex objects were potent in attracting early fixations, in support of a saliency map model of scene inspection. In the second experiment viewers were set the task of detecting the presence of a low saliency target, and the effect of a high saliency distractor was negligible, supporting a model in which the saliency map can be built with cognitive influences that override low-level visual features.

Address for correspondence:
Geoffrey Underwood, School of Psychology, University of Nottingham,
Nottingham NG7 2RD, UK. *E-mail:* geoff.underwood@nottingham.ac.uk
Acknowledgements:
A preliminary report of this investigation was presented at the meeting of Experimental Psychology Society at the London meeting in January 2005. We are grateful to Laurent Itti for use of software for the measurement of visual saliency, to Jan Theeuwes, Andrew Hollingworth, Marc Brysbaert, Peter De Graef and two anonymous reviewers for their comments on previous drafts of this paper, and to the Nuffield Foundation for award URB/01455/G that enabled us to complete this project.

What attracts attention when we inspect photographs of natural scenes, and how does the task influence eye guidance during scanning? By recording eye fixations on specific objects shown in a photograph of a natural scene we can determine the objects that attract attention, and the order in which they are fixated can be taken as an indication of their saliency. Unlike the eye movements made while reading text (Rayner, 1998) or while reading musical notation (Gilman & Underwood, 2003), where the structure of the array requires a sequence of fixations in a well-defined order, the order of fixations made to scenes is not prescribed by the need to generate a sequenced output. Scenes can, in principle, be inspected with fixations made in a number of sequences, although regularities can be observed. These regularities have given rise to a theoretical account of the process by which we extract information from pictures that links fixation behaviour with the development of a cognitive representation of the scene under scrutiny. This representation is built through a "saliency map" of the low-level, visually informative regions of the display (Koch & Ullman, 1985; Findlay & Walker, 1999; Henderson, Weeks & Hollingworth, 1999; Itti & Koch, 2000; Parkhurst, Law, & Niebur, 2002), and the purpose of the experiments described here is to determine the effects of visual salience in two experimental tasks that require the viewer's attention when they first see a picture of a natural scene.

The initial studies of Buswell (1935) and Yarbus (1967) demonstrated concentrations of fixations upon foreground objects such as people and salient objects, leading to the suggestion that it is the information within areas of a picture that attracts our attention. As with a reader's eye being attracted to the informative parts of a word (Hyönä, Niemi & Underwood, 1989; Underwood, Clews & Everatt, 1990), the question is how did the viewer know that this was a location of high information value before it had been fixated. The implication is that parafoveal vision can deliver sufficiently detailed information for a preliminary analysis of content to be conducted and used in the programming of eye movements. Mackworth and Morandi (1967) have reported evidence of the more frequent inspection of regions of pictures that were independently rated as being more informative. Fixations upon informative regions were made within 2 sec of seeing the photographs, suggesting that an analysis of the meaningful elements of a picture can be made during early visual processing. Antes (1974) reported a similar result, with the first saccade frequently being to an informative region. A different approach was used by Loftus and Mackworth (1978), who observed fixations made to line drawings containing incongruous objects (e.g., a tractor in an underwater scene, or an octopus in a farmyard). Incongruous objects were fixated earlier than the same objects appearing in a congruous drawing (e.g., a tractor in a farmyard, or an octopus in an underwater scene). Fixation of an incongruous object would occur immediately after the first fixation, again suggesting early analysis of the meaningful configuration of the scene. This result has been challenged, however, with De Graef, Christiaens and d'Ydewalle (1990) and Henderson et al. (1999) failing to find an effect of semantic

incongruity on early fixation behaviour (see also, De Graef, 1998; Henderson & Hollingworth, 1998).

Current accounts of eye guidance during scene comprehension place semantic analysis late in the process. A saliency map is first developed using low-level features including colour, contour density, and luminance, and with weightings assigned to regions that are then allocated attention depending upon the current weighting (Koch & Ullman, 1989; Itti & Koch, 2000; Parkhurst et al., 2002). This principle has been developed in a formal model of saliency that can convert a picture of a natural scene and, after separating the image into three channels (to identify colour, intensity and orientation), develop a map that represents the major areas of visual conspicuity. The map highlights areas of change that would enable a viewer to discriminate one scene from another, and the Parkhurst et al. (2002) eye-tracking study found a high correlation between fixations and saliency. This correlation supports the view that low-level visual features determine the selection of initial fixation locations, but Parkhurst et al. set their viewers a free-viewing task in which no encoding for a memory test was used, and in which no decisions were made as part of the task. There was little or no cause for cognitive processes to influence the selection of fixation locations in their experiment.

In the Findlay and Walker (1999) model, what they term the "salience map" is a spatiotopic representation of weightings that can be thought of as troughs and peaks that will influence the decision about the location of the next fixation. In visual search tasks, saccade trajectories are determined by the peak that is currently dominant, and in this version of the model it is the 'where' decision that is controlled by the saliency map. Additionally, this model acknowledges the influence of top-down cognitive influences, and provides an explanatory framework for the appearance of fixation patterns that are not predicted by low-level visual descriptions of the scene.

The Henderson et al. (1999) "saliency map framework" also builds a representation of the scene in which regions of interest are identified with an analysis of low-level visual information. The first fixation is attracted to the region with the greatest weighting, and the duration of that fixation is determined by the complexity of processing, and this is taken to include both perceptual and semantic processing. At this point the map starts to incorporate meaningful information about the gist of the scene, and objects may be identified. Upon completion of processing, the saliency weighting of the inspected information is reduced, and attention is re-allocated to the next region with high saliency. The early fixations on a scene are thereby determined primarily by visual processes, and only after fixation can the saliency map of a scene be represented with semantic weightings. Fixation patterns made on pictures accompanied by text support the view that the gist can be determined during the first few fixations, and that an extended search is not necessary in order to build a representation of the main features of the picture. In a range of tasks viewers characteristically move their gaze from the graphical component after just two or three fixations, in order to read the accompanying text in a mixed display (Carroll, Young & Guertin, 1992; Rayner, Rotello, Stewart, Keir & Duffy, 2001; Underwood, Jebbett & Roberts, 2004). The early departure from scene to text suggests that sufficient information can be extracted during this time to develop a representation that can be used for reference when reading the sentence.

Henderson et al. (1999) supported their model with two experiments in which viewers' eye movements were recorded while they inspected line drawings of familiar scenes. Scenes sometimes contained incongruous objects such as a drawing of microscope in a bar-room, and the congruous equivalent of this would be the same scene with a cocktail glass replacing the microscope. In the first experiment viewers were free to inspect the scene in preparation for a memory test, and in the second they searched for a target object in order to make a present/absent decision. The target varied from trial to trial, and in the search task it was specified by a verbal label prior to onset of the display. The two tasks were used as a check of the hypothesis that viewers may be motivated to find incongruity only when the task requires a specific search. There was no evidence of early fixation of incongruent objects in either experiment. In the memory experiment there were about 10 fixations prior to object fixation, in contrast with less than four fixations in the search task, but in neither case was there a difference in the patterns with congruent or incongruent objects. Fixation durations on the drawings in the memory experiment did vary, with incongruent objects attracting longer fixations. Once discovered, they also attracted more fixations than their congruent counterparts.

The saliency map theory makes specific predictions about the inspection of scenes for different purposes. Fixations should be attracted to regions of high salience – high contour density, high contrast, high luminance, colour change and other low-level visual features – and eye fixations should be attracted regardless of the semantic content of the picture. They should also be attracted regardless of the relevance of the semantic salience of the region to the specific purpose of inspection.

The two experiments here test the predictions of the saliency map hypothesis using the same tasks as used by Henderson et al. (1999), but with photographs of natural scenes. Specifically, the scenes contained two objects that differed in their saliency, as determined by the Itti and Koch (2000) algorithm. In the memory experiment there was no differentiation between these objects in terms of their significance for the task to be performed, but in the search experiment the viewer was set the task of determining whether or not a low saliency target object was present. Both experiments asked whether fixations would be attracted primarily to the object with higher saliency, as predicted by the saliency map hypothesis, and whether the purpose of inspection can provide a cognitive override that makes visual saliency of secondary importance.

Experiment 1: Inspection for encoding

As with Experiment 1 of Henderson et al. (1999), viewers were here shown a series of pictures, and were instructed that they were to inspect them in preparation for a memory test. We used photographs rather than line drawings. Each photograph was taken in an office environment and showed a collection of objects on an office surface (desktop etc), with two objects of interest being positioned either side of centre. One of these two objects had high visual saliency – it was the most conspicuous object in the picture - and the other had lower saliency. The experimental measures of eye fixations were used to determine which object received primary attention.

Figure 1
An example of the output of the Itti & Koch (2000) saliency program (stopped after identification of the three most salient objects in the scene) superimposed on one of the pictures used in Experiments 1 and 2. The pictures were shown in colour in both experiments. The most salient part of the picture is the food jar, followed by the lemon, and then the bunch of keys. In this example the high saliency object (food jar) is located 3 deg from centre, and the low saliency object (lemon) is 6 deg from centre. In the search task used in Experiment 2 the low saliency object (the piece of fruit) was used as the target, and the high saliency object became a distractor.

Method

Participants. Twenty students (aged 18-26 years) each received £5 for their participation in this experiment, and all had normal or corrected-to-normal vision.

Stimuli and Screening. Digital photographs of office scenes were displayed on a computer monitor (1024 x 768 pixels) at a distance of 60 cm from the seated participant, generating a colour image that subtended 11.6 deg by 15.4 deg from this viewing position. There were 48 office scenes containing objects in a cluttered environment, and of central interest was a desk, shelf or table containing a number of objects. Eight further

pictures were prepared to give participants practice with the memory task. The photographs always included a low saliency object and sometimes contained a very conspicuous object with high saliency, with eight possible combinations of these objects in different positions. Other objects such as books and computer equipment were also visible. The low saliency object was a piece of fruit, although this had no significance for the participants in this experiment. Different fruits were used in each picture. The high saliency objects were of similar size and colour as the fruits, but with greater contour density. These objects were coffee mugs with decorations and commercial packages with patterns. Neither of these objects had text visible, in case items of text might attract attention. The two objects of interest were located along the horizontal meridian at either 3 deg or 6 deg from the centre of the picture. These values were selected on the basis that Henderson et al. (1999) found that the saccadic amplitude when inspecting pictures was approximately 3 deg. A near object was placed so that it could be fixated with just one saccade, by this arrangement. The six possible combinations of objects therefore had the low saliency object at 3 deg or at 6 deg from centre, and a high saliency object that was absent, or at 3 deg, or at 6 deg. There were eight pictures in each of these six possible arrangements. The two objects appeared equally often on either side of the picture, and when both objects were present they appeared on opposite sides of the scene.

A saliency map of each picture was determined using software provided by Laurent Itti, and that is described by Itti and Koch (2000). This map identifies the saliency, or conspicuity, of objects in the picture, on the basis of variations in orientation, intensity and colour. The program's default weightings were employed, to avoid prioritisation of any of these three dimensions of saliency. The output from the program used here consists of a version of the original picture with a series of circles identifying the most conspicuous objects in rank order of their saliency weightings. In the pictures used in this experiment one of the objects of interest (a manufactured object) was selected and placed so that it would be the most salient object in each scene. The second object of interest (a piece of fruit) had lower saliency. Pictures that did not meet this criterion were replaced until this object was identified by the program as the object of greatest saliency in all 48 pictures. An example of a picture used in the experiment is shown in Figure 1, together with the first three outputs from the program that identifies the objects of highest saliency.

Apparatus. Eye movements were recorded using a SensoMotoric Instruments (SMI) EyeLink system that was also used to collect keyboard responses to each display. The eye-tracker was head-mounted, and recordings taken from the participants' right eye every 4 msec. The spatial accuracy of the eyetracker is better than 0.5 deg. Head position was recorded remotely, but to minimise movements and to ensure a constant viewing distance a chin-rest was also used.

Procedure. Participants were initially calibrated for recording eye position with the SMI eye-tracker, and were instructed that the study concerned scene memory. They were told that their eye movements would be monitored while they inspected photographs of office scenes in preparation for a recognition test. This recognition test was never actually administered during the experiment, and was only used during a practice session that showed an additional set of eight pictures with four two-choice tests. Our focus of interest was the eye fixations during initial identification of the two critical objects, not the memory of the picture. The presentation of each picture was preceded by a drift correction marker, to confirm central fixation of the screen. Each participant saw all 48 pictures presented in a unique randomised order, and each picture was shown until the participant pressed a computer keyboard key.

Table 1
Fixation of the low saliency object in Experiment 1, as a function of the presence and location of a high saliency distractor. [Standard deviations are in parentheses.]

	Low saliency object at 3 deg			Low saliency object at 6 deg		
High saliency object:	None	3 deg	6 deg	None	3 deg	6 deg
No. of fixations prior to fixation of the low saliency object	3.86 [4.11]	4.43 [3.14]	6.11 [2.67]	2.76 [2.28]	5.17 [3.50]	5.17 [2.65]
Duration of 1st gaze (msec) of the low saliency object	483 [176]	501 [136]	456 [141]	473 [225]	357 [139]	440 [225]
Total inspection of picture (msec)	5781 [2675]	6155 [2978]	6084 [2919]	5187 [2260]	6211 [3641]	6032 [2683]

Results and Discussion

Only eye fixations on the two objects of interest were analysed. Fruits were the relatively low saliency objects here, and the other object had the highest saliency of all objects in the scene. Only fixations in excess of 50 msec were scored. The three measures taken were: the number of fixations made prior to fixation of each object; the duration of the first gaze on each object, and the duration of inspection of the scene up to the point when the participant indicated that they were ready to see the next picture. The means of these measures were calculated and used for the within-group ANOVA tests. These means are presented in Table 1 (inspection of the low saliency object) and Table 2 (inspection of the

high saliency object). The high saliency objects were fixated on 84.8% of trials in Experiment 1.

Table 2
Fixation of the high saliency object in Experiment 1, as a function of the location of both objects of interest. [Standard deviations are in parentheses.]

	Low saliency object at 3 deg		Low saliency object at 6 deg	
High saliency object:	3 deg	6 deg	3 deg	6 deg
No. of fixations prior to fixation of the high saliency object	3.79 [1.37]	6.24 [2.73]	2.23 [1.34]	4.82 [3.14]
Duration of 1st gaze (msec) on the high saliency object	447 [201]	507 [147]	542 [193]	578 [300]

Number of fixations prior to fixation of the objects. The number of fixations made between onset of the picture and first fixation of the object of interest was taken as a indication of how long it took the participants to find the low saliency object in the presence or absence of a more conspicuous object, and to determine which object was most effective in attracting attention. The fixation at the centre of the screen, made during the onset of the display of the picture, was included in this count. Three ANOVAs were performed, to determine the effects of a high saliency object upon the first fixation of the other object, upon the first fixation of the high saliency object itself, and to compare the time preceding fixation of the two objects for those pictures that showed both of them.

The analysis of fixations prior to inspection of the low saliency object had two factors – eccentricity of the low saliency object and eccentricity of the high saliency object. The analysis found that eccentricity of the low saliency object was not a reliable factor ($F(1, 19) = 1.29$; MSe = 4.32) but that there was an effect of the eccentricity of the high saliency object upon fixation of the fruit ($F(2, 38) = 8.52$; MSe = 6.51; $p < 0.001$). Scheffé tests were used to inspect the influence of the three levels of eccentricity of the high saliency object. Fewer intermediary fixations were made when there was no object than when the high saliency object was at 3 deg ($p < 0.05$) or at 6 deg ($p < 0.01$). There was no interaction between the eccentricities of the two objects ($F(1, 19) = 2.56$; MSe = 4.03).

A second analysis was performed on the number of fixations made prior to the first fixation on the high saliency object, using just those trials when it had been presented. There was a main effect of the eccentricity of the high saliency object, with earlier fixation of the nearest objects ($F(1, 19) = 20.66$; MSe = 6.15; $p < 0.001$). There was also an effect of the location of the other object upon the time taken to fixate the high saliency object ($F(1, 19) = 23.08$; MSe = 1.92; $p < 0.001$), with more fixations prior to

inspection when the low saliency object was in a near location. The interaction was not reliable ($F < 1$).

A third analysis was performed using the data from those trials where both objects were shown. The data used for this analysis are summarised in Tables 1 and 2. The analysis had three factors – eccentricity of the low saliency object, eccentricity of the high saliency object, and the object of inspection (low/high saliency). There was a main effect of the object of inspection ($F(1, 19) = 8.01$; MSe = 6.01; $p < 0.05$), with fewer fixations prior to inspection of the conspicuous object. There was also a main effect of the eccentricity of the low saliency object ($F(1, 19) = 6.38$; MSe = 3.85; $p < 0.05$), with fewer fixations prior to fixating an object at 3 deg than one at 6 deg. This main effect of the position of the low saliency object is seen as an influence on the fixation of both objects, but an interaction indicated a greater effect upon the fixation of the high saliency object ($F(1, 19) = 9.93$; MSe = 1.88; $p < 0.01$). This was inspected with an analysis of simple main effects. The position of the low saliency object had an effect upon the number of fixations preceding inspection of the conspicuous object ($F(1, 19) = 11.17$; MSe = 3.85; $p < 0.01$), with earlier inspection of these high saliency objects when the low saliency object at 6 deg than for when it was at 3 deg, but there were similar numbers of fixations prior to fixation of the low saliency object whatever the eccentricity of the low saliency object ($F < 1$).

A main effect of eccentricity of the high saliency object was not reliable ($F(1, 19) = 3.15$; 6.39), but this eccentricity factor was also involved in an interaction ($F(1, 19) = 18.13$; MSe = 5.27; $p < 0.001$). The analysis of simple main effects indicated an effect of the position of the conspicuous object upon fixation of that object ($F(1, 19) = 15.93$; MSe = 6.39; $p < 0.001$), but no effect of eccentricity of the low saliency object on the number of fixations prior to fixation of the conspicuous object ($F < 1$).

Duration of first gaze on the objects. The duration of the first inspection of an object or word is often an indication of the difficulty of processing (see, for example, Rayner, 1998; Henderson et al., 1999; Underwood et al., 2004), and so gaze duration may also be indicative of the difficulty of recognising the objects in the pictures shown here. Gaze is defined as the total of all fixations made on the object prior to a fixation upon another object. If there is only one fixation, the most frequent case here, then first gaze duration is equivalent to first fixation duration. If the viewer looks first at one part of an object and then re-fixates on another part, before looking at another object, then gaze is the sum of the two fixations.

A three-factor ANOVA was applied to the first gaze durations using eccentricity of the low saliency object, eccentricity of the high saliency object, and which object was fixated. The object being inspected was a reliable main effect ($F(1, 19) = 4.67$; MSe = 0.0546; $p < 0.05$), with shorter gazes on the low saliency object than on the high saliency object. No other main effects were reliable. Neither eccentricity of the high saliency object ($F(1, 19) = 1.91$; MSe = 0.0228), nor eccentricity of the low saliency object ($F < 1$) modified gaze duration. An interaction between object inspected and eccentricity of the low saliency object ($F(1, 19) = 13.08$; MSe = 0.0203; $p < 0.01$) was inspected with an analysis of simple main effects. This indicated that there were shorter gazes on low saliency objects that were presented at 6 deg than on those presented at 3 deg ($F(1, 19) =$

4.54; MSe = 0.0281; $p < 0.05$), but that there were longer gazes on high saliency objects accompanied by low saliency objects presented at 6 deg than at 3 deg ($F(1, 19) = 4.93$; MSe = 0.0281; $p < 0.05$).

Duration of inspection of the picture. The display remained on screen until the participant pressed a key to indicate that they had encoded the scene and were ready to proceed to the next picture. These inspection times are shown in Table 1, and were submitted to a two-factor ANOVA. Eccentricity of the low saliency object did not influence the total inspection time ($F(1, 19) = 2.09$; MSe = 0.5542), but eccentricity of the high saliency object was effective ($F(2, 38) = 4.87$; MSe = 1.1433; $p < 0.05$). Scheffé comparisons indicated only one effect, with longer inspections when the conspicuous object was close to the centre of the picture than when this object was absent ($p < 0.05$).

Summary of Experiment 1

When inspecting a picture in preparation for a memory test, the relatively featureless fruits received less attention overall than the more conspicuous mugs, drinks cans and other manufactured products. This is a reflection of the relative contour densities and variations in colour and intensity in those parts of the picture, and is also indicated in the durations of the first fixations on these two objects. The object with greater visual complexity attracted fixations earlier and for longer, as predicted by the saliency map hypothesis.

The essential results from these analyses are as follows, with a clear influence of the visually complex objects upon the inspection of the more uniform fruits. When the conspicuous object was absent, fixation on the fruit took fewer saccadic movements than when it was present. Conspicuous objects were fixated earlier than less salient objects. The two objects of interest were always presented on opposite sides of the screen, and so inspection of an object furthest from centre will necessarily incur a greater cost upon the subsequent fixation of the other object. Although the high saliency objects were fixated earlier than the other objects, as predicted by the saliency model, they were not the first objects shown in the pictures that attracted attention. As the means in Table 2 indicate, conspicuous objects that were placed at 3 deg from centre screen were fixated after two saccadic movements, and those at 6 deg were fixated after 4.53 saccadic movements on average. These were the most salient objects in the scene, but this did not ensure that they would be the first locations to be inspected.

There were a number of influences of each object on the other in these analyses, as well as differences in the inspection of the two objects. When a conspicuous object was present in the picture, it took longer to fixate the visually simple fruit. This effect of an object competing for the viewer's attention was reciprocated by an effect of a simple fruit upon a complex mug or package by the high saliency object receiving a shorter inspection when the piece of fruit was located near to the centre of the picture. The low saliency object may have acted as a distractor here, by attracting attention away from the more conspicuous object. The long-standing question of fixation and gaze durations is re-visited here: is a long duration indicative of more difficult processing of the object under

scrutiny, or is it an indication of a longer preview prior to re-fixation? Given the difficulty of fixating one object while attending to another in a different location, it is more likely that long gazes indicate longer processing of the object under scrutiny, and this view is supported by the appearance of shorter gazes on the visually more simple objects. When there are two objects in proximity they will influence each other, however, prompting shorter gazes and increased delays prior to fixation. In the present experiment, these influences were mutual.

Experiment 2: Inspection for target search

Henderson et al. (1999) reported that their target objects received early fixations independently of their meaning in their memory experiment. Incongruous objects were fixated no earlier than their congruous equivalents, suggesting that the saliency map did not represent semantic information early enough to attract primary fixations. An alternative explanation considered by Henderson et al. is that in a memory task the viewers had no motivation to attempt to identify areas of incongruity. They used a search task in their second experiment, to prompt viewers to look for these targets. Although incongruous objects were fixated no earlier than congruous objects, both types were fixated earlier than in their memory experiment, confirming the effectiveness of task instructions. The same approach is used here, with viewers instructed to determine whether the scene contained a piece of fruit. The strong version of the saliency map hypothesis would require that complex objects, such as the high saliency objects used in Experiment 1 and used again here as distractors, should continue to attract attention early in the inspection of the scene. Initial fixations should be directed to the high saliency objects. A weaker version of this hypothesis suggests that the effects of salience can be modified by task demands. The weak version of the hypothesis, supported by Henderson et al., suggests that if the task is possible using low-level visual features, and discrimination between targets and distractors in the present experiment was indeed possible using visual complexity, then a saliency map will be able to take task demands into account. If this task information can be used to override the dominance of high contour density and colour variation seen in Experiment 1, then early fixations on the low saliency target object should be observed.

The same pictures were shown here as were used in Experiment 1, with the low saliency object designated as a target by instructing viewers to say whether or not there was a piece of fruit in the scene. The high saliency objects in Experiment 1 were distractors in this search task.

Participants. Twenty students (aged 18-26 years) participated in this experiment, all of whom had normal or corrected-to-normal vision. None of the participants had taken part in Experiment 1.

Apparatus and stimuli. The same digital photographs used in Experiment 1 were also used here, and presented using the same computer screen while eye movements were recorded with the SMI eye-tracker. Twenty-one additional pictures were used, containing a complex distractor but no target fruit, so that the search task sometimes required a

negative response. When a target was present, it occurred equally often in the four possible positions. There were also 10 practice pictures used, none of which appeared during the experiment sequence. Two keys on the computer keyboard were used to collect the decision responses which were recorded on the computer controlling the experiment.

Procedure. The only change from Experiment 1 was in the instructions given to the participants, which now required them to say whether the picture showed a piece of fruit. The computer keyboard was placed in front of the participants, with the instruction to press either a *yes* or a *no* key according to whether there was a piece of fruit shown in the picture. The pictures were shown in an order randomised for each participant, and the display was terminated when the response key was pressed.

Table 3
Fixation of the low saliency target object in Experiment 2, as a function of the presence and location of a high saliency distractor. [Standard deviations are in parentheses.]

	Target at 3 deg			Target at 6 deg		
Distractor:	None	3 deg	6 deg	None	3 deg	6 deg
No. of fixations prior to target fixation	1.16 [0.25]	1.41 [0.27]	1.28 [0.19]	1.37 [0.31]	1.67 [0.34]	1.49 [0.39]
Duration of 1st gaze (msec)	327 [134]	305 [117]	280 [99]	293 [138]	301 [141]	325 [144]
Total inspection of picture (msec)	627 [202]	677 [198]	691 [200]	610 [163]	607 [159]	648 [146]

Results and Discussion

The search task was performed with an accuracy greater than 95% and so all trials where a target was present were included in the analyses. Distractors were fixated on 20.4% of occasions in this experiment, and so analyses corresponding to those done in Experiment 1 were not always possible. The measures recorded in Experiment 1 were used again here. The means for the number of fixations prior to this first fixation, duration of the first fixation and total inspection time, are presented in Table 3.

Fixations prior to fixation of the objects. The numbers of fixations prior to the first fixation on the target were entered into a two-factor ANOVA with target eccentricity and distractor eccentricity as factors. A two-factor ANOVA applied to the number of

fixations prior to first fixation of the target found a reliable effect of target eccentricity ($F(1, 19) = 22.08$; MSe = 0.07; $p < 0.001$), with fewer fixations for closer targets. The main effect of distractor eccentricity was reliable ($F(2, 38) = 15.07$; MSe = 0.05; $p < 0.001$), and Scheffé comparisons indicated that when there was no distractor there was earlier fixation upon the target than when there was a near distractor ($p < 0.001$), and also a difference between near and far distractors ($p < 0.05$), with near distracters delaying fixation on the targets to a greater extent. A distractor at 6 deg had a similar effect to no distractor, and near distractors had the greatest effect in delaying target fixation. There were insufficient numbers of fixations on distractors to compare the time to first fixation of target and distractor.

Duration of first gaze on the target. The durations of the first fixations on the targets were submitted to an ANOVA with the two factors of eccentricity of target and of distractor. There was no main effect of target eccentricity ($F < 1$) or of distractor eccentricity ($F < 1$), but there was an interaction between the two factors ($F(2, 38) = 6.46$; MSe = 0.0025; $p < 0.01$), and this was further inspected with an analysis of simple main effects. The effect of a distractor was only effective upon the gaze duration on a near target ($F(2, 38) = 3.91$; MSe = 0.0029; $p < 0.05$), with no effect upon far targets ($F(2, 38) = 1.89$; MSe = 0.0029). The only paired comparison that was reliable by Scheffé comparisons was that, with near targets, far distractors prompted shorter gazes than absent distractors ($p < 0.05$).

Inspection of the pictures. The analysis of the total amount of attention given to the picture indicated an effect of target eccentricity ($F(1, 19) = 6.84$; MSe = 0.0082; $p < 0.05$), with longer inspections when the target was nearer to centre screen. There was also a main effect of distractor eccentricity ($F(2, 38) = 3.89$; MSe = 0.0067; $p < 0.05$), and Scheffé comparisons indicated that the only effect was for longer inspections with far distractors than for pictures with no distractor ($p < 0.05$).

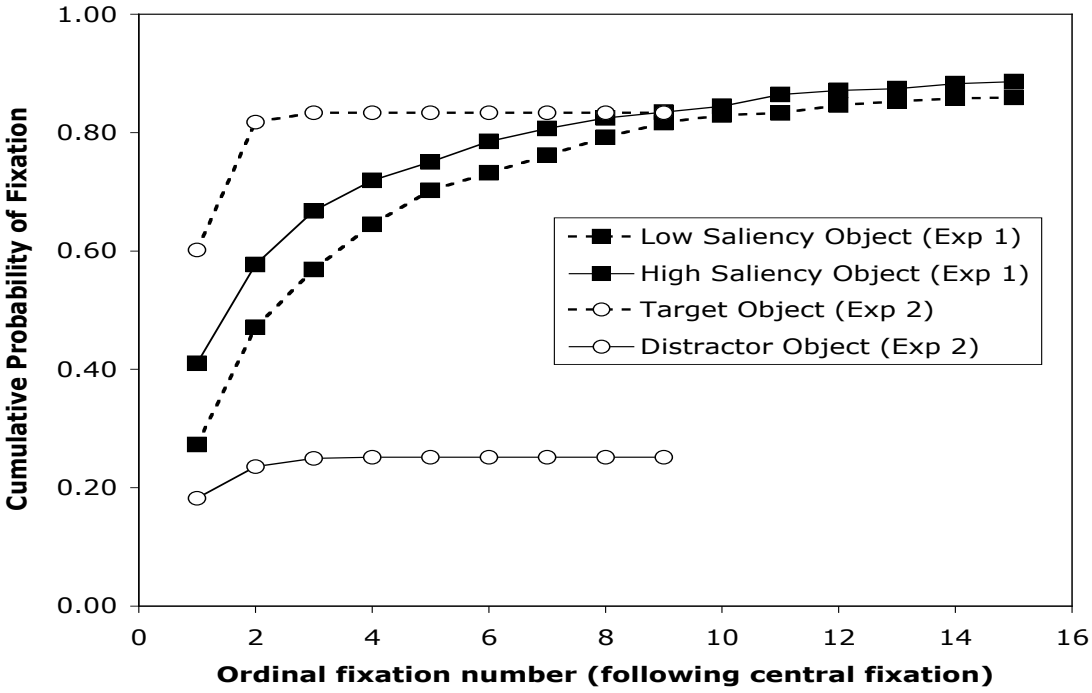
Summary of Experiment 2

When searching for a pre-specified object viewers were only slightly influenced by a high saliency distractor, and much less than when freely scanning the picture in preparation for a memory test. The distractor was fixated on only one fifth of all trials. Distractors had marginal effects on target fixation, and these effects were not seen to the same extent as in the memory experiment. The only reliable effect of attentional capture by the high saliency object was that when there was a distractor present, then the target was fixated later than when it was absent, and that near distractors were more disruptive than those furthest from the target. Once attention had been captured by the target, the distractor had a minimal influence on the gazes on targets. In contrast to Experiment 1, gazes were shorter and less variable. There was one effect of a distractor on the gaze on far targets, possibly a result of a conspicuous object continuing to attract attention during the decision process after the target had been found but before full identification.

These analyses suggest that visual saliency was of secondary importance when the task is to search for a well-specified object. To compare the inspection of the pictures in

the two experiments where different tasks were used, a further series of ANOVAs was conducted on the data from target fixations.

Figure 2
Cumulative probability of fixation of the two objects of interest in Experiments 1 and 2, as a function the number of fixations made from the onset of the display. Note that the low saliency objects in Experiment 1 were used as the target objects in Experiment 2, and that the high saliency objects became the distractors in Experiment 2. In the search experiment the cumulative probability of fixating the target or the distractor does not increase after the fourth fixation, and the plots are continued here only for comparison with those from the memory experiment.



Comparison of Experiments 1 and 2

Where the numbers of measures allowed, direct comparisons were possible between the means from the memory and search experiments. As the distractors were infrequently fixated in the search experiment, these comparisons were restricted to the attention given to the targets (low saliency objects), and the effects of distractors (high saliency objects) on their fixation. The comparisons used the data that are summarised in Tables 1 and 3. These analyses were intended to highlight the differences attributable to task differences between the memory task and the search task.

A three-factor mixed design ANOVA was applied to the number of fixations prior to fixation of the low saliency object, with experimental task (memory vs. search) as the between-groups factor and eccentricity of the low saliency object (near/far) and of the high saliency object (near/far/absent) as the within-groups factors. There were fewer fixations prior to inspection in Experiment 2 than in Experiment 1 ($F(1, 38) = 36.57$; MSe = 16.69; $p < 0.001$), and eccentricity of the low saliency object was again not reliable ($F < 1$). Eccentricity of the high saliency object was a reliable factor ($F(2, 76) = 9.73$; MSe = 3.28; $p < 0.001$), although an interaction with experiments ($F(2, 76) = 7.42$; MSe = 3.28; $p < 0.05$) that was inspected with simple main effects indicated that the effect of location of the high saliency object was restricted to the memory task ($F(2, 76) = 16.91$; MSe = 3.28; $p < 0.001$), and was ineffective in the search task ($F < 1$).

The number of fixations prior to first fixation of each object of interest is indicated in Figure 2, which shows the cumulative probability of fixating each object as a function of the ordinal fixation number. In Experiment 1 fixation of the low saliency object occurred after 5.22 fixations on the picture, and this was later than fixation on the high saliency object (4.12 fixations). In the search task however, fixation of the low saliency target occurs after 1.23 fixations and on those occasions when the high saliency distractor was fixated, it was inspected after 2.21 fixations.

Finally, a comparison was made between experiments using the measure of total inspection time. Experimental task had a substantial influence ($F(1, 38) = 73.05$; MSe = 22.7697; $p < 0.001$), with much longer inspections in the memory task than in the search task. In addition there was a main effect of eccentricity of the high saliency object ($F(2, 76) = 5.35$; MSe = 0.5750; $p < 0.01$), with longer inspections of the picture only if this object was present. The position of the high saliency object interacted with experimental task however ($F(1, 38) = 4.37$; MSe =0.5750; $p < 0.05$), and Scheffé comparisons found that it was only in the memory task that there were briefer inspections on pictures with no high saliency object, relative to pictures showing either a near or a far conspicuous object ($p < 0.01$ and $p < 0.001$, respectively). There were no differences between near and far distractors, and no effects of distractor position in the search task.

A striking difference between the two experiments concerns the number of fixations on the high saliency distractor. In Experiment 1, where pictures were encoded in preparation for a memory test, this object was fixated on 84.5% of trials, but in Experiment 2, where viewers were searching for a piece of fruit, it was fixated on only 20.4% of trials. This object had the greatest conspicuity of all objects shown in each picture, but the cognitive demands of the search task could be seen to override this high visual saliency.

General Discussion

The dominance of the high saliency object in the memory experiment was not seen in the search task. This requires rejection of the strong version of the saliency map hypothesis in which low-level visual features should determine eye movements during the early inspection of a scene regardless of cognitive demands. The most salient object in the picture attracted eye fixations, but this attentional capture was seen only in the memory experiment. When the same pictures were shown for a different purpose, then fixation

patterns changed accordingly, and the visually most salient object could be disregarded. The fixation of objects was influenced by the presence of other objects – high and low saliency objects were seen to influence the inspection of each other in the memory task. As the task changed, so did the scanning behaviour. High saliency objects were non-targets in the search task, and the distractors were inspected infrequently. When viewers were looking specifically for a piece of fruit the distractors had minimal effect on inspection of this target.

This pattern of results supports a version of the saliency map hypothesis in which task demands can override the saliency map. The saliency weights of objects in a scene can be modified by the need to identify a specific object, and the attractiveness of conspicuous objects thereby minimised by cognitive saliency. The inspection of an object with a smooth surface was disrupted by the presence of a more discriminable object with high contour density and contrasting colour and intensity, but there was more disruption during a free inspection task in preparation for a memory test than in a directed search task. This is taken as a demonstration of the potency of visual complexity in attracting eye fixations, and supports the saliency map hypothesis of scene perception, but only for those cases where the scene is under free inspection (Koch & Ullman, 1985; Findlay & Walker, 1999; Henderson, Weeks & Hollingworth, 1999; Itti & Koch, 2000; Parkhurst, Law & Niebur, 2002). When inspecting a picture with a specific purpose, as in the search task, visual saliency is secondary to cognitive demand. The ability of top-down cognitive processes to influence the use of the saliency map is recognised in developments of the model that take task demands into account (Navalpakkam & Itti, 2005).

The saliency map hypothesis suggests that early fixations on a scene are guided by low-level visual features, such as the high contour density and by colour and intensity variation of the objects shown, and only when a saliency map has been built using these features can semantic information be extracted and used to represent the scene in terms of the meaningful configuration of objects depicted. The second experiment qualifies this support for the saliency map hypothesis, in that a change in task demands from general encoding to directed search had the effect of reducing the potency of highly salient but task-irrelevant distractors. These distractors were inspected on less than a quarter of the occasions that they were available, and they had minimal influence on the inspection of the targets. The search experiment can only support a weaker version of the hypothesis, in which high-level demands set by the viewer's intentions can influence the saliency weightings used in the development of the representation of the scene (Henderson et al., 1999). Visual saliency is effective in free-viewing, but can be overridden by cognitive demands. When searching a picture for an object distinguishable with one low-level feature, other low-level features can be disregarded. If selecting one particular piece of fruit from a plate, for example a strawberry, the weak saliency map hypothesis suggests that we would not be distracted by an orange, or by green apples and pears, or by objects of similar size that can be discriminated by purely visual features. This is the version of the saliency map hypothesis consistent with both experiments here, and the only version supported by the directed search experiment.

The cognitive override of visual saliency as the task changes is consistent with the Findlay and Walker (1999) version of the model, which acknowledges top-down influences in eye guidance with three processes. Their process of "spatial selection" can

modify the saliency weights and thereby enable inspection of a low saliency region that may contain an object of interest such as a target in a search task. A process of "search selection" can also override the visual saliency map by directing fixations to objects that share visual features with the target object, and the third process of "intrinsic saliency" uses the viewers own knowledge of the scene to guide fixations to probable target locations. In the search task here, for example, the viewer would not look for a piece of fruit floating in mid-air, but only at supportive surfaces. When violations are introduced in contrived pictures, objects that violate these expectations are recognised more slowly than those that comply with the viewer's knowledge of the properties of objects (Biederman, Mezzanotte & Rabinowitz, 1982).

The processes of cognitive override that were proposed by Findlay & Walker have now been refined in a model of eye guidance proposed by Tatler, Baddeley and Gilchrist (2005). Four competing models were evaluated with data from a picture memory task, in which viewers' eye fixations were recorded as they inspected photographs of scenes in preparation for cued-recall questions. The four models vary in their use of the representation built with low-level visual saliency values. The "salience divergence" model suggests that visual and cognitive saliency weightings change over time as information is extracted from the scene. The bottom-up component is initially dominant in this model, until the objects in the scene are recognised. At this point eye guidance is determined by the semantics of the scene, as proposed by Henderson et al. (1999) and Parkhurst et al. (2002). Tatler et al. rejected this model on the basis that there was no variation in the saliency values of fixated and non-fixated locations in their memory task. The "salience rank" model rank orders locations in the scene on the basis of their saliency weights, and these ranks are used to guide fixations from one location to another. This is the model closest to the proposals of Itti and Koch (2000). The fixation patterns in the memory task did not follow the saliency ranks, and so this model was also rejected. The third model is called the "random selection with distance weighting" model, and follows the proposal by Melcher and Kowler (2001) that visual and cognitive saliency have little influence on the selection of target locations relative to the influence of distance between objects. This model predicts variability between viewers inspecting the same scene, whereas in the Tatler et al. experiment there was consistency in the locations of early fixations. The final possibility considered was the "strategic divergence" model. In this model the visual saliency map does not change during the period of inspection, but the influence of cognitive strategic factors does change. The visual saliency map provides a frame of reference for inspection of the scene, but the task demands and the individual knowledge and interests of the viewer will determine the objects that are to be fixated. This model was supported by the early consistency of fixation locations in the Tatler et al. experiment, and by the increased divergence between viewers over the course of inspection.

The results from the present experiments are consistent with the Tatler et al. (2005) strategic divergence model, in that the saliency weightings predicted the early fixations on a scene in free inspection and because cognitive influences were seen to modify these fixation patterns. The results can be used to reject the salience divergence model, because there was no evidence of a change of top-down and bottom-up influences in the search task. There was no influence of visual saliency in the early selection of

saccadic targets in this task. The salience rank model predicted the early fixation locations in the memory task, but not when cognitive demands intervened in the search task. Finally, the random/distance model can be rejected by the appearance of predictable and consistent inspection patterns in the inspection of the two objects of interest in each picture. This model may find more support from studies in which more extensive sets of locations are inspected, of course, and so this must be a qualified rejection based on the analysis of a restricted set of locations. The strategic divergence model accounts for variations between cognitive tasks and between individuals, and Findlay and Walker's (1999) three processes of spatial selection, search selection, and intrinsic saliency provide mechanisms whereby cognitive processes can influence the eye guidance system.

The effectiveness of changing task instructions on search behaviour is evident in the comparison between the memory and search experiments reported by Henderson et al. (1999). When searching for a specified object, viewers made fewer fixations to the target, saccadic amplitudes to the targets were larger, and fixation durations were slightly shorter than when encoding the whole picture for a memory task. Similar changes in fixation behaviour can be induced by task demands that are implicit rather than declared in the instructions. Underwood et al. (2004) had two groups of participants inspect the same photographs, with a task of deciding whether an accompanying sentence was an accurate description of events in the picture. One group saw the picture before the sentence, and therefore had to encode the whole scene in preparation for a sentence that could have referred to any part of it, while the other group read the sentence first, and therefore knew what to look for when the picture was displayed. The first group had a general encoding task when viewing the pictures, and the second group had a directed search task. These differences were again reflected in the fixation patterns, with many more fixations in the encoding task, although there was no difference in fixation duration, in contrast with the viewers in Henderson et al.'s (1999) experiments and in contrast with the data from the present experiments. The first fixation on the target item was considerably longer in the memory task than in the search task here. The difference between the implicit and explicit forms of the search task, whereby shorter fixations on the target are found only with an explicit target detection task, are likely to be a product of the difference between the time taken by encoding processes as against that taken by decision processes. In both tasks the object must be identified, but when it and its relationships to other objects are encoded then fixations are longer than when a decision is taken about its status as a target object.

The potency of low-level visual factors in determining the order of inspection of objects in a complex scene is especially notable in the present memory experiment. Visually complex distractors were fixated earlier and for longer than the relatively simple fruits that were to serve as the targets in the search experiment. General inspection is guided by a saliency map that is based upon featural complexity and fixations are attracted to regions that are most distinguished from the surrounding background. This conclusion is consistent with other views of the search process. On the basis of a series of experiments in which simple line and texture targets were detected against uniform backgrounds, Geisler and Chou (1995) concluded that most of the variance in search time can be predicted from discrimination functions based upon low-level visual factors. Background complexity is critical to the discrimination process, and using natural scenes rather than uniform backgrounds, Wolfe, Oliva, Horowitz, Butcher and Bompas (2002)

have reported that it is only when background objects become barely distinguishable from the target object that the serial stage of item checking affected. Candidate targets can be segmented from the background preattentively in parallel in their Guided Search Model, with attention then moving to the locations occupied by the candidates. The saliency map hypothesis would also regard the initial segmentation stage as using low-level visual features to develop the saliency map, and which is then consulted by the saccade generator in determining which objects to inspect in turn.

Low-level visual saliency was effective in capturing attention in the encoding task but not in the search task. An unresolved issue concerns the ways in which the saliency of an object works to capture attention and attract an eye fixation, and an alternative model dispenses with the notion of saliency maps altogether. Our low saliency objects were pieces of fruit, and possessed relatively uniform colouring with few internal edges. In contrast, the high saliency objects were multi-coloured manufactured products with distinct internal and external edges. The Itti and Koch (2000) program selected these objects as being the most conspicuous. However, it may have been that the conspicuous objects attracted attention in the encoding task as a result of visual discontinuities, as predicted by the saliency map model, or alternatively because they were too complex to be identified with peripheral vision. The model of visual capture tested here is based on the saliency map model, in which attention is successively attracted to the next highest peak on the map. The alternative model of capture by visual complexity suggests that the objects in a scene are recognised in parallel and with the extensive use of peripheral vision, and in which attention is attracted to objects that are too complex for analysis with peripheral vision. Reports of the rapid recognition of the gist of scene support the parallel identification of individual objects without fixation, and fixation may then be necessary for objects with complex detail. These objects also gain longer fixations than visually simpler equivalents. In the encoding experiment the pieces of fruit may have been fixated later because they could be identified using peripheral vision, or because they are visually inconspicuous. This model of scene perception does not rely upon a saliency map at all, and guides attention and eye movements to objects on the basis that they are too complex for recognition without fixation. These alternatives are not resolved by the present study.

When inspecting pictures with the intention of encoding them in preparation for a memory test, eye fixations are attracted by informationally-rich, complex objects, but when specifically instructed to search for a target that possesses relatively low contour density and uniform colouring and intensity, then the effects of a visually more complex distractor are minimised. The saliency map that is used to guide saccadic movements around the scene is determined initially with low-level visual features, but task demands can moderate the influence of high saliency values of complex objects. Eye movements are then determined by cognitive saliency, and visual saliency can be neglected.

References

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology, 103,* 62-70.

Biederman, I., Mezzanotte, R. J. & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14,* 143-177.

Buswell, G. T. (1935). *How people look at pictures: A study of the psychology of perception in art.* Chicago: University of Chicago Press.

Carroll, P.J., Young, J.R. & Guertin, M.S. (1992). Visual analysis of cartoons: A view from the far side. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 444-461). New York: Springer-Verlag.

De Graef, P. (1998). Prefixational object perception in scenes: Objects popping out of schemas. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 313-336). Oxford: Elsevier.

De Graef, P., Christiaens, D. & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research, 52,* 317-329.

Findlay, J. M. & Walker, R. (1999). A model of saccade generation base on parallel processing and competitive inhibition. *Behavioral and Brain Sciences, 4*, 661-721.

Geisler, W. S. & Chou, K.-L. (1995). Separation of low-level and high-level factors in complex tasks; Visual search. *Psychological Review, 102,* 356-378.

Gilman, E. & Underwood, G. (2003). Restricting the field of view to investigate the perceptual spans of pianists. *Visual Cognition, 10,* 201-232.

Henderson, J.M. & Hollingworth, A. (1999). Eye movements during scene viewing; An overview. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 269-293). Oxford: Elsevier.

Henderson, J.M., Weeks, P.A. & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 210-228.

Hyönä, J., Niemi, P. & Underwood, G. (1989). Reading long words embedded in sentences: informativeness of word parts affects eye movements. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 142-152.

Itti, L. & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489-1506.

Koch, C. & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4,* 219-227.

Loftus, G. R. & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 565-572.

Mackworth, N. H. & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics, 2,* 547-552.

Melcher, D. & Kowler, E. (2001). Visual scene memory and the guidance of saccadic eye movements. *Vision Research, 41,* 3597-3611.

Navalpakkam, V. & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45,* 205-231.

Parkhurst, D., Law, K. & Niebur, E. (2002). Modelling the role of salience in the allocation of overt visual attention. *Vision Research, 42,* 107-123.
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 372-422.
Rayner, K., Rotello, C.M., Stewart, A.J., Keir, J. & Duffy, S.A. (2001). Integrating text and pictorial information: Eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied, 7,* 219-226.
Tatler, B. W., Baddeley, R. J. & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research, 45,* 643-659.
Underwood, G., Clews, S. & Everatt, J. (1990). How do readers know where to look next? Local information distributions influence eye fixations. *Quarterly Journal of Experimental Psychology, 42A,* 39-65.
Underwood, G., Jebbett, L. & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search. *Quarterly Journal of Experimental Psychology, 57A,* 165-182.
Wolfe, J. M., Oliva, A., Horowitz, T. S., Butcher, S. J. & Bompas, A. (2002). Segmentation of objects from backgrounds in visual search tasks. *Vision Research, 42,* 2985-3004.
Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum Press.