

## Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search

Geoffrey Underwood, Lorraine Jebbett, and Katharine Roberts

*University of Nottingham, Nottingham, UK*

When we see combinations of text and graphics, such as photographs and their captions in printed media, how do we compare the information in the two components? Two experiments used a sentence–picture verification task in which statements about photographs of natural scenes were read in order to make a true/false decision about the validity of the sentence, and in which eye movements were recorded. In Experiment 1 the sentence and the picture were presented concurrently, and objects and words could be inspected in any order. In Experiment 2 the two components were presented one after the other, either picture first or sentence first. Fixation durations on pictures were characteristically longer than those on sentences in both experiments, and fixations on sentences varied according to whether they were being encoded as abstract propositions or as coreferents of objects depicted in a previously inspected picture. The decision time data present a difficulty for existing models of sentence verification tasks, with an inconsistent pattern of differences between *true* and *false* trials.

When we read literature such as newspapers and magazines, as well as textbooks and scientific journals, we frequently need to integrate text in the form of a caption or legend that accompanies a picture, graph, or diagram. Relative to the attention given to reading research, or picture perception, the extraction of information from combinations of sentences and pictures has been largely neglected. The observation of eye movements while reading and inspecting pictures has generated a good understanding of the processes involved (for reviews, see Kennedy, Radach, Heller, & Pynte, 2000; Rayner, 1998; Underwood, 1998), but few studies have recorded eye fixations while participants integrate the text that accompanies a picture.

---

Correspondence should be addressed to Geoffrey Underwood, School of Psychology, University of Nottingham, Nottingham, NG7 2RD, UK. Email: [geoff.underwood@nottingham.ac.uk](mailto:geoff.underwood@nottingham.ac.uk)

This work was supported by an EU 5th Framework Programme i-Eye project (contract IST-1999-11883). Thanks are due to Kari-Jouko Räihä (project coordinator) and other members of the i-Eye team at the University of Tampere (Finland), Conexor (Finland), SensoMotoric Instruments (Germany), and GIUNTI Interactive Labs (Italy) for their comments on the studies described here. James Cupit provided help in writing experimental control programs for data collection. Martin Pickering, Keith Rayner, Simon Liversedge, Peter Chapman, David Crundall, and Anna Green provided helpful comments on a previous draft. A preliminary version of this paper was presented at the joint meeting of the EPS and the BVP/SBP, held in Leuven in April 2002.

How do viewers compare these two sources of information? Is it important to encode the arrangement of objects in the picture first, or to read the text first? How does the order of presentation influence the pattern of eye movements used when extracting information from a legend with an accompanying picture, graph, or diagram?

The question of order of inspection and its effects upon fixation patterns has been addressed previously by Carroll, Young, and Guertin (1992), who presented line drawings and captions from a range of Gary Larson's "Far Side" cartoons. Viewers were presented with the caption and the cartoon, with the task being to understand the cartoon. In their first experiment the two components were presented together, as they would appear in a newspaper, and the second experiment controlled the order of presentation. When the two components were presented at the same time there was some variability between individuals, but with a general tendency to look briefly at the cartoon, with as few as three fixations, then to look at the caption in order to read it carefully, and then to refixate the drawing, this time making more fixations, but focusing upon the objects and characters mentioned in the caption. There was little evidence of repeated relocation of the direction of gaze between the drawing and the caption, and viewers tended to read the entire caption before returning to the drawing. Hegarty (1992a, 1992b) reported a similar result using captions that described the movements of pulleys in mechanical diagrams accompanying the text. Eye-movement recordings indicated that the diagrams were inspected only after the text had been read, although viewers sometimes stopped reading at the end of a clause in order to inspect part of the diagram (Hegarty, 1992a, 1992b). In almost all cases however, the viewers appeared to be attempting to understand the text before inspecting its referent in the diagram. After establishing the preferred viewing order of text and diagram with concurrent presentations, the second of the Carroll et al. experiments then investigated the effects of presenting the two parts of the caption/drawing combination separately. Viewers saw first the drawing and then the caption, or they saw the caption followed by the drawing. Different order effects appeared in the inspection of the caption and the drawing. When the text was presented first, it attracted more fixations, resulting in longer overall reading time, than when it was presented second. Fixation durations on the captions remained constant. For the drawings, the total inspection time did not vary whether it appeared first or second, but fixation durations were longer when the drawing appeared after the caption.

Longer fixation durations are generally indicative of more extensive processing (Rayner, 1998), and reflect a difficulty such as when a reader encounters a word of low frequency (e.g., Rayner & Duffy, 1986) or a word that is contextually implausible (e.g., Ehrlich & Rayner, 1981). A similar pattern emerges when viewers inspect objects in pictures, with incongruous objects attracting longer fixations than those that are natural components of the scene (Henderson, Weeks, & Hollingworth, 1999; Loftus & Mackworth, 1978). There are exceptions to the general relationship between fixation duration and difficulty, with drivers showing shorter fixation durations when encountering more demanding roads (Crundall & Underwood, 1998) and when watching video recordings of demanding road scenes taken from a moving vehicle (Chapman & Underwood, 1998; Underwood, Chapman, Bowden, & Crundall, 2002). This exception to the rule is a product of a need to fixate an increased number of objects in an immediate hazardous environment displayed dynamically, when events occur at a pace that is not determined by the observer. Long fixations are an indicator of uneventful scenes rather than difficult processing when driving along empty rural roads or when

watching video recordings of these featureless scenes. However, when inspection of the display is self-paced, as in reading, then fixation duration is a reliable indicator of processing difficulty.

The report of longer fixations when cartoon drawings were inspected after the cartoon caption (Carroll et al., 1992) indicates that the drawing is more difficult to understand when it is placed in the context of a caption, and it must then be integrated with the idea presented by the already-encoded caption. Longer fixation durations here are indicative of the viewers computing the humorous intent, including resolution of the incongruity designed by the cartoonist.

Combinations of text and graphics often appear in a more integrated format in newspaper and magazine advertisements, with the text superimposed over part of a picture. However, even for these more composite presentations Rayner, Rotello, Stewart, Keir, and Duffy (2001) have reported patterns of eye fixations that generally agree with the results from Carroll et al. (1992). Adults looked at advertisements for a number of products, with instructions that they should view them as if they were interested in buying one of them. Their scanpaths indicated initial fixation upon the picture part of the display, followed after approximately three fixations upon the accompanying text, and then refixation of the picture. Similarly, viewers tended not to make repeated saccadic movements between the text and the pictures. In both studies, fixations on the picture part of the display were longer than those on the captions. There were differences between the studies in the saccadic amplitudes reported, with Rayner et al. finding longer saccades on the pictures than on the text, and Carroll et al. reporting shorter saccades on the line drawings. This difference is most probably due to task differences and to the differences in the pictures used in the two studies. The advertisements used by Rayner et al. contained some large objects whereas the cartoon line drawings used by Carroll et al. were fine-grained and informationally dense.

The approach taken here, to the question of how we extract information from combination or multi-format displays, was to have participants respond to a statement about a picture. The written statement required inspection of the picture and was either true or false, and so this is a version of the sentence verification task. The present experiments therefore provide a test of the generality of models of the sentence verification task for the case of pictures of scenes more detailed than those that have been used in the past (Carpenter & Just, 1975; Clark & Chase, 1972). The sentence verification paradigm was considered to be an ideal instrument because it requires inspection of both the picture and the sentence without the introduction of information that would be exclusive to either. The sentence verification task is also well understood and allows the derivation of specific predictions about performance in the picture-sentence version used here.

Simple diagrams containing geometric shapes (Clark & Chase, 1972) or a display of dots (Gough, 1965) may be encoded completely in preparation for a simple sentence verification task without difficulty, even when shown before the appearance of sentences such as "The star is above the cross" or "The dots are red". Generating an abstract propositional model such as that proposed by Clark and Chase and by Carpenter and Just (1975) would require few operations if the picture contained only two geometric shapes or a few coloured dots. With a photograph of a natural scene, however, there would be many features that could be the subject of the verification sentence, and presenting the picture before the sentence would then be expected to result in a need for a more extensive encoding strategy.

A robust finding from verification tasks is that sentences requiring denial are more difficult than those requiring a confirmation response, and this is the pattern of results predicted for the picture–sentence version of the task. Early versions of the task required the verification of simple declarative sentences that were true or false, such as “Seven is an even number” (Wason & Jones, 1963) or “A canary is a bird” (Collins & Quillian, 1969). Slower responses to *false* trials are also reported in tasks requiring verification of a statement about a simple picture that accompanied it. Judgements about the locations of coloured shapes (Wason, 1959), about the colour of printed dots (Gough, 1965), about the relative number of one set of dots against a set in a different colour (Just & Carpenter, 1971), and about the juxtaposition of two simple geometric shapes (Clark & Chase, 1972), are each performed with false instances requiring more processing time than true statements.

Explanations of this effect, such as the Constituent Comparison Model (Carpenter & Just, 1975), suggest that the performance of sentence verification tasks requires that the sentence should be internally represented as an ordered set of abstract propositions formed from the constituents, that the picture should also be represented as a set of constituent propositions, and that the two sets of constituents should then be compared serially. Sentences calling for falsification require longer responses, according to this model, because an additional comparison is required when a mismatch is detected. This model, supported by an extensive literature, predicts that when the sentence in the following experiments is an incorrect description of the contents of the picture, then the response will be slower than when it provides a true description (Clark & Chase, 1972; Gough, 1965; Just & Carpenter, 1971; Wason, 1959). By tracking the viewer’s eye movements during inspection of the sentence and picture, the present experiments identify the components of decision making. Longer fixations are an indication of more difficult processing, and increased durations on the sentence or on the picture would indicate the source of the falsification effect.

The falsification effect appeared to be robust during the early development of models of sentence verification, but more recently reports have appeared that question its universal applicability. Feeney, Holo, Liversedge, Findlay, and Metcalfe (2000) reported an advantage for true or matching statements about bar graphs (“A is greater than B”) over false or mismatching statements, although the effect was moderated by the order of presentation of terms in the sentence. Deciding whether “A is less than B” against a bar graph showing the opposite to be true was faster and more accurate than deciding about “B is less than A” (true) and about “B is greater than A” (false). False statements can receive faster decisions than true statements. Goolkasian (1996, 2000) has also presented evidence from a version of the sentence verification task that questions the generality of the falsification effect. In her experiments statements were verified against sets of simple objects (pictures of geometric shapes and similarly simple objects), and whereas the falsification effect was robust for verbatim verifications (“There are seven red squares” against a mixed display of red, yellow, and blue squares), differences between *true* and *false* statements were minimal for verifications requiring an inference (“Yellow squares are least likely”). Falsification appears now to require more than the serial comparison of abstract propositions in which an additional processing stage is required relative to the verification of true statements. It may be that with some displays the comparison between picture and sentence is not made using the same relational term as that in the sentence, especially when the verification requires reasoning. The present experiments further question the generality of the falsification effect during the inspection of statements

about natural scenes. The models of sentence verification described by Clark and Chase (1972) and by Carpenter and Just (1975) allow us to predict slower decisions when the sentences do not accurately describe the events shown in the picture, relative to those cases where sentence and picture are in agreement.

When reading the sentence first, we can then focus upon the specific features of the picture that have been foregrounded, but when we inspect the picture first there are more features to encode and more relationships to infer. Inspection of the picture must be more extensive with this order of viewing, to identify all of the possible aspects of the picture that might be the subject of the sentence. Accordingly, if we first read a sentence that describes a scene or a relationship between two depicted objects, then we can dedicate our attention to the specific part of the picture containing the critical objects. If we inspect the picture first, then the whole of the scene must be captured, in order to cover any of the objects or relationships that might be mentioned in the text. In the development of their saliency map model of scene inspection, Henderson et al. (1999) used different tasks with the same line drawings to encourage the use of different inspection strategies. In their first experiment, viewers were presented with a memory task, in which they were to look at the drawings in preparation for a recognition memory task. In their second experiment the task was to search for a specified object. The first fixation on the target object was sooner in the search experiment than it was in the memory experiment, demonstrating the effectiveness of task instructions. This difference is used in the present experiments, but within the sentence verification framework. The equivalent of the memory experiment is predicted when the picture precedes the sentence, because with this ordering the viewer cannot know which aspect of the picture will be questioned. Viewers should encode as much of the picture as possible, in contrast with our equivalent of the Henderson et al. search experiment in which the sentence is presented first. With a sentence–picture ordering the focus of interest is defined prior to presentation of the picture, and so inspection should be directed to the relevant parts of the picture to the exclusion of detail not mentioned in the sentence.

Two experiments were conducted to address the question of how the order of acquisition influences our fixation behaviour, using pictures of natural scenes and sentences that declared some relationship between aspects of the scene depicted. The task was to decide whether the sentence was accurate or inaccurate. In the first experiment a picture and a sentence appeared together, with the viewer free to inspect the two components of the display in any order. In the second experiment the order of inspection was imposed upon the viewer, with either the picture or the sentence appearing first and the second component appearing only when the viewer indicated that encoding was complete. The first experiment was more representative of the combinations of text and picture that appear in printed media. In the second experiment the order of presentation was controlled, to determine the effect of having to encode either the picture or the sentence before the other part of the display appears. This is essentially the same procedure as that used by Carroll et al. (1992), in which the first experiment presented the two components in the same display, and the second experiment presented the two components sequentially. Whereas Carroll et al. had their viewers attempt to understand a cartoon, the sentence–picture verification task here invites a simple prediction of *true* statements gaining faster decision than *false* statements. Carroll et al., Hegarty (1992a, 1992b), and Rayner et al. (2001) have reported a fixation scanpath during the inspection of mixed-format displays in which a small number of fixations on the picture (line drawing or advertisement graphic) is

followed by a reading of the text. A brief inspection of the picture may be sufficient to develop the low-level saliency map of the main features of the picture (Henderson et al., 1999), and this is the pattern of fixations predicted for the photographs of natural scenes used here. If the saliency map of a picture containing rich detail cannot be developed with a few fixations, then we should observe more fixations on the picture prior to inspection of the sentence.

## EXPERIMENT 1 Concurrent displays

Participants viewed pictures and sentences together in this experiment, with the sentence directly below the picture on each trial. In addition to the overall decision time determined by a keyboard response, eye movements were recorded to provide measures of the number of fixations on each part of the screen and the durations of those fixations. Because the pictures used in the experiment were roadway scenes photographed from the perspective of a car driver, we ensured that all participants would have some familiarity with the scenes by requiring that they were all car drivers.

### Method

#### *Participants*

A total of 24 university students were paid to participate in this experiment. All had normal or corrected-to-normal eyesight, and all held UK driving licences.

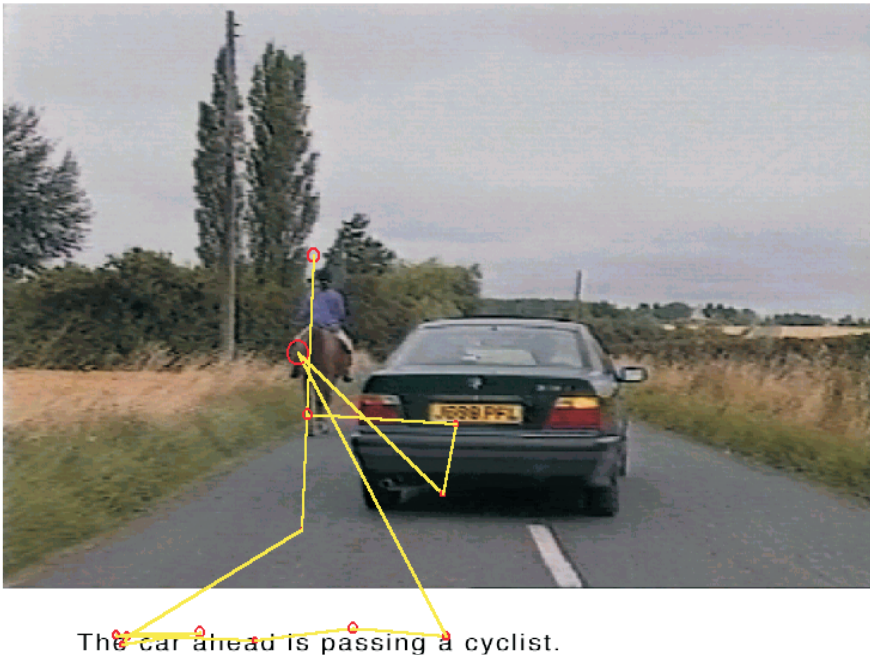
#### *Stimuli and design*

The 32 stimuli presented to each participant were a combination of a digitized colour photograph of a roadway scene plus a sentence displayed directly below the picture on a computer monitor. The monitor screen subtended angles of 28 degrees (horizontally) and 22 degrees (vertically) at the viewing distance of the participants. The pictures subtended angles of 18 degrees and 14 degrees, respectively, at the required viewing distance of 77 cm, and the sentences were presented with approximately five character spaces per degree. An example of a picture/sentence display used in this experiment is presented in Figure 1.

Each photograph was taken from a car driver's perspective, on a range of roads and with a characteristic range of other road users and roadway objects in view. The sentence appearing below the picture described a plausible action (for example, "The car ahead is passing a cyclist"), or an arrangement of objects depicted in the scene (for example, "There is a roadside snack bar on the left"). Each sentence was a simple active declarative statement that could be verified by inspection of the picture. Two versions of each sentence were constructed for each picture, one true and one false. Half of the participants were presented with a picture accompanied by a true statement, and the other half saw the same picture accompanied by a false statement. Each participant saw 16 true statements and 16 false statements. The sentences were each between 4 and 12 words in length, and each sentence appeared on the monitor as a single line of text.

#### *Apparatus*

The SensoMotoric Instruments (SMI) EyeLink system was used to display the stimuli and collect eye movement recordings and keyboard responses. This system sampled eye positions every 4 ms, with



**Figure 1.** An example of a picture and sentence from Experiment 1. The circles represent fixations, with duration indicated by diameter. The first fixation in this example is slightly to the left of the centre of the picture.

an average error in determining the location of fixation of less than 0.5 degrees. The system monitored head position remotely, to allow participants to view the screen without head restraint, and in addition a chin-rest was used to minimize head movements. The screen was refreshed at 85 Hz. The eye fixation and head location recorders were mounted on a headband.

### *Procedure*

Participants were first fitted with the headband-mounted eye-tracker, and the system was calibrated using the SMI EyeLink program. They were then instructed to judge the accuracy of each statement about the accompanying picture by pressing either a key marked “true” or one marked “false”. Participants were randomly allocated to one of two groups, with true and false versions of the sentences alternated between groups. Each display started with a 1-s fixation marker in the centre of the screen, followed immediately by the picture and sentence. These stimuli remained on the screen until one of the two response keys was pressed. After a 3-s interval the next display was presented. Within each set of stimuli the order of presentation was randomized.

### **Results and discussion**

The dependent measures that were recorded were the overall keyboard response time, the total inspection time on each part of the display (the sum of all fixations), the type of response made (true or false), the number of fixations made on the picture and those made on the sentence, and the durations of those fixations. Fixations less than 60 ms were excluded from all analyses. Accuracy was high (>95%), and so only correct responses were analysed. Each

analysis was performed both with participants as the random variable ( $t_1$  and  $F_1$ ) and with items as the random variable ( $t_2$  and  $F_2$ ).

Table 1 presents all of the data from Experiment 1 with the exception of the keyboard response time to each display. This is the time between onset of the display and the keyboard response to indicate that the sentence was true or false. *False* responses (4.97 s,  $SD = 2.46$  s) took longer than *true* responses (4.27 s,  $SD = 1.91$  s),  $t_1(23) = 2.50, p < .05; t_2(15) = 3.81, p < .01$ . This result is consistent with previous work using a range of sentence verification tasks, and it supports the Clark and Chase (1972) and Carpenter and Just (1975) models.

The remaining data were analysed with repeated measures analysis of variance (ANOVA), with sentence validity (true/false) and part of display (picture/sentence) as the two factors.

The measure of inspection time presented in Table 1 was the total duration of all fixations on the sentence. These combined total durations of fixations on the picture and the sentence are less than the overall keyboard decision time because they exclude the time taken by saccadic eye movements, blinks, and fixations away from the screen. For example, if a participant looked down at the keyboard to select the appropriate response key, this would be aggregated with the decision time but not with the inspection time. The analysis of inspection time revealed a main effect of validity,  $F_1(1, 23) = 5.87, p < .05; F_2(1, 15) = 8.58, p < .05$ , with longer inspections for false (2.11 s) rather than true statements (1.81 s). There was an unreliable effect of the part of the display inspected,  $F_1(1, 23) = 2.53; F_2(1, 15) = 1.08$ , and the interaction was also unreliable,  $F_1(1, 23) = 3.07, p < .05; F_2(1, 15) = 1.51$ .

Table 1 also shows the mean number of fixations made on each part of the display when sentences were true and when they were false. A two-factor ANOVA found that false displays (9.22 fixations) received more fixations than true displays (7.98 fixations),  $F_1(1, 23) = 7.99, p < .01; F_2(1, 15) = 9.38, p < .01$ , and that the sentences received more fixations (9.43) than pictures (7.77),  $F_1(1, 23) = 6.18, p < .05; F_2(1, 15) = 3.77, p < .05$ . There was no interaction.

The mean fixation durations are also shown in Table 1, and an ANOVA found no effect of validity,  $F_1$  and  $F_2 < 1$ , but a main effect of display,  $F_1(1, 23) = 135.18, p < .001; F_2(1, 15) = 69.97, p < .001$ . Fixations on the pictures (265 ms) were longer than those made when reading the sentences (185 ms). There was no interaction.

Experiment 1 presented the sentence and picture together and required participants to decide whether or not the sentence described the objects shown in the picture. The overall decision was longer for *false* trials than for *true* trials, as predicted on the basis of previous

TABLE 1  
Eye fixation means for inspection of the two components of concurrent displays in  
Experiment 1

	<i>Sentences</i>				<i>Pictures</i>			
	<i>True</i>		<i>False</i>		<i>True</i>		<i>False</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Inspection time <sup>a</sup>	1.78	0.81	1.82	0.63	1.85	0.99	2.40	1.76
Number of fixations	9.18	3.41	9.68	2.82	6.79	3.01	8.76	5.61
Fixation duration <sup>b</sup>	186	23.8	185	21.3	266	49.6	265	32.9

<sup>a</sup>In s. <sup>b</sup>In ms.

experiments using more abstract displays, and the result is qualified by the analysis of eye fixations. The total inspection time on the picture, and separately on the sentence, again showed an advantage for *true* trials.

The characteristic inspection pattern or scanpath started with a fixation near to the centre of the picture. A fixation marker at the start of each trial requested that the participant look at a part of the computer monitor where the picture would appear. Within three fixations, typically, their eyes would saccade to the sentence, and they then read the sentence completely before inspecting the picture and made the decision immediately following this second visit to the picture. This is the pattern shown in Figure 1. The mean number of fixations on the picture prior to a fixation on the sentence was 2.63 ( $SD = 0.77$ ) for *true* trials and 2.88 ( $SD = 0.74$ ) for *false* trials. These means did not differ reliably,  $t_1(23) = 1.03$ ;  $t_2(15) < 1$ . This confirms the pattern of scanpaths reported by Carroll et al. (1992), Hegarty (1992a, 1992b), and Rayner et al. (2001), who found inspection of the textual component of the display starting after a small number of fixations on the graphics. In the present experiment, some participants moved their eyes between sentence and picture a number of times, but the decision about the validity of the sentence was usually made not while reading the sentence but while viewing the picture. There were also more fixations on *false* trials than on *true* trials. Sentences attracted more fixations than pictures, but the durations of these fixations were shorter than those on pictures. Previous studies have also found shorter fixations while reading sentences than while inspecting pictures (Carroll et al., 1992; Rayner et al., 2001).

Given that the first fixation on the picture was induced at the start of each trial, the rapid change of gaze location from picture to sentence raises the question of whether there was any information extraction from the picture prior to reading the sentence. The experiment does not answer this question, but Sanocki and Epstein (1997) suggest that perception of a scene can be facilitated by prior presentation of a priming scene that makes the layout available early. The first few fixations may therefore have served a purpose in generating an initial representation of the overall spatial layout of the picture. After reading the sentence this spatial representation would be used to guide the inspection of the referents identified in the sentence as relevant to the verification decision.

## EXPERIMENT 2

### Successive displays

To further separate the processing of the picture and sentence components of the displays, participants in this experiment saw each component individually. Half the participants read the sentence first, and on pressing a response key to indicate comprehension, the sentence was removed and the picture displayed. The other participants first looked at the picture, which was removed when they pressed a key to indicate that they had encoded it, and the sentence then appeared. On the basis of the Henderson et al. (1999) experiments, in which viewers either encoded a line drawing in preparation for a recognition test or searched for a specific object, we anticipated that when seeing the picture first they should attempt to encode as many features from the picture as possible, and when seeing the sentence first they should direct their attention to the events foregrounded by the sentence.

## Method

### *Participants*

A total of 48 university students were paid to participate in this experiment. All had normal or correct-to-normal eyesight, all held UK driving licences, and none had participated in Experiment 1.

### *Stimuli, design, apparatus, and procedure*

The same pictures and sentences as those used in Experiment 1 were again used here, with the components of the display separated. Sentences and pictures were prepared as separate displays within a trial. Participants saw either a sentence followed by picture, or a picture followed by a sentence, and each participant was presented with 32 of these pairs of stimuli. Each participant saw only sentence–picture pairs, or picture–sentence pairs, with 16 pairs in which the sentence accurately described the events in the picture (*true* trials) and 16 pairs in which a word or phrase was replaced with a plausible alternative, to create *false* trials. Participants were randomly allocated to the sentence–picture group or to the picture–sentence group. Within these groups there was a further division of participants, to enable rotation of materials. A *true* version of each sentence was presented to half the participants in each group, and a *false* version to the other half.

The eye-tracking apparatus used in Experiment 1 was also used here. After calibration, each participant was instructed that they would be shown a series of sentences and pictures, and that the task was to judge whether the sentence accurately described the contents of the picture. The first part of the display (a sentence for half the participants and a picture for the other half) remained on screen until the space bar was pressed. Prior to the sentence a fixation marker appeared at the beginning of the line, and prior to the picture a fixation mark appeared in the centre of the screen. When the space bar was pressed after the first part of each trial, the second part of the display (a picture or a sentence, respectively) appeared and remained on the screen until one of two response keys was pressed. After a 3-s interval the next trial started.

## Results and discussion

Each trial consisted of two components: a sentence and a picture. The order of presentation was consistent within each of the two groups of participants. For half the participants the first component was a sentence, and the second component was a picture, and for the other participants the order of presentation of the components was reversed. Measures of total inspection time, number of fixations, and duration of fixation were taken, in addition to the time taken before pressing the space bar in response to the first component, and the time taken to decide whether the sentence accurately described the picture and the accuracy of the response. Participants' responses to the second component were 83.6% accurate ( $SD = 7.72$ ) when the sentence appeared first, and 79.2% ( $SD = 15.2$ ) when the sentence appeared second. The difference between these two groups of participants was not reliable,  $t_1(46) = 1.28$ ;  $t_2(15) < 1$ , and no further analysis was performed on the accuracy data. In the subsequent analyses only data from trials were used if the response had been accurate.

Table 2 presents the means for the four measures of overall performance. The means were entered into a series of mixed-design ANOVAS, each with three factors (order of presentation, validity of sentence, and part of display). In each case ANOVAS were performed to obtain  $F_1$  (by participants) and  $F_2$  (by items) ratios.

TABLE 2  
Response time and eye fixation means for inspection of the two components of successive displays in Experiment 2

	<i>Sentence first</i>								<i>Picture first</i>							
	<i>Sentences</i>				<i>Pictures</i>				<i>Sentences</i>				<i>Pictures</i>			
	<i>True</i>		<i>False</i>		<i>True</i>		<i>False</i>		<i>True</i>		<i>False</i>		<i>True</i>		<i>False</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Response time	2.71	0.68	2.76	0.82	2.44	0.99	2.69	1.14	3.31	0.87	3.39	1.12	5.05	2.24	4.76	2.20
Inspection time <sup>a</sup>	1.99	0.57	2.04	0.72	1.75	0.84	1.97	0.98	2.37	0.75	2.45	0.97	4.05	2.01	3.80	1.94
Number of fixations	9.36	2.67	9.54	3.26	6.25	2.80	7.26	4.03	9.27	2.43	9.54	3.12	14.72	7.05	14.15	7.01
Fixation duration <sup>b</sup>	213	21.9	214	20.9	281	42.1	282	40.8	253	42.0	257	38.1	276	33.3	273	30.3

<sup>a</sup>In s. <sup>b</sup>In ms.

The keyboard response times to each of the two displays that comprised a trial were the responses made to terminate each part of the display. For the first part of the display this was the bar press that terminated the first component and started displaying the second component. For the second part of the display this was the keyboard response that indicated the decision as to whether the statement was *true* or *false*. The ANOVA indicated that the order of presentation was a reliable main effect,  $F_1(1, 46) = 19.32, p < .001$ ;  $F_2(1, 15) = 7.50, p < .05$ , with faster responses when the sentence was shown first (2.65 s) than when it was shown second (4.13 s). There was no effect of validity of the sentence,  $F_1 < 1$  and  $F_2(1, 15) = 1.60$ . The effect of display type was reliable,  $F_1(1, 46) = 14.04, p < .001$ ;  $F_2(1, 15) = 16.79, p < .001$ , with sentences (3.04 s) receiving faster responses than pictures (3.74 s). These main effects were moderated by two interactions. Order of presentation interacted with type of display,  $F_1(1, 46) = 21.79, p < .001$ ;  $F_2(1, 15) = 7.59, p < .05$ , and an analysis of simple main effects indicated no difference in the keyboard times to sentences and pictures when the sentence appeared first, both  $F_1$  and  $F_2 < 1$ , but when the picture appeared first it elicited a longer response time than the sentence,  $F_1(1, 92) = 37.26, p < .001$ ;  $F_2(1, 15) = 37.84, p < .001$ . There was also a three-way interaction that was not confirmed by the items analysis,  $F_1(1, 46) = 4.36, p < .05$ ;  $F_2(1, 15) = 2.27$ .

The results of the ANOVA performed on the total inspection times exactly matched those from the analysis of keyboard response times, with main effects of presentation order,  $F_1(1, 46) = 17.14, p < .001$ ;  $F_2(1, 15) = 7.19, p < .05$ , and display type,  $F_1(1, 46) = 16.55, p < .001$ ;  $F_2(1, 15) = 26.18, p < .001$ , and no effect of validity,  $F_1 < 1$  and  $F_2 < 1$ . There was again an interaction between order and display type,  $F_1(1, 46) = 25.26, p < .001$ ;  $F_2(1, 15) = 5.78, p < .05$ . An analysis of simple main effects indicated that there was more attention given to pictures when they appeared first,  $F_1(1, 46) = 41.36, p < .001$ ;  $F_2(1, 15) = 53.12, p < .001$ , and no difference between sentences and pictures when the sentences appeared first,  $F_1$  and  $F_2 < 1$ . There was also a three-way interaction that was not confirmed by the items analysis,  $F_1(1, 46) = 4.85, p < .05$ ;  $F_2 < 1$ .

The analysis of the number of fixations made to each component of each trial revealed a main effect of presentation order,  $F_1(1, 46) = 12.46, p < .001$ ;  $F_2(1, 15) = 5.29, p < .05$ , with more fixations made when the picture was presented first (11.92) than when it was presented second (8.10). Neither display type,  $F_1(1, 46) = 3.91$ ;  $F_2(1, 15) = 1.03$ , nor validity,  $F_1 < 1$ ;  $F_2(1, 15) = 3.80$ , appeared as main effects, but display type interacted with presentation order,  $F_1(1, 46) = 42.89, p < .001$ ;  $F_2(1, 15) = 20.50, p < .001$ . An analysis of simple main effects found an effect of order for pictures,  $F_1(1, 46) = 38.84, p < .001$ ;  $F_2(1, 15) = 10.42, p < .01$ , but not for sentences,  $F_1$  and  $F_2 < 1$ . When pictures were presented first they attracted more fixations than sentences,  $F_1(1, 46) = 36.36, p < .001$ ;  $F_2(1, 15) = 40.15, p < .001$ , but when sentences were presented first they attracted more fixations than pictures,  $F_1(1, 46) = 10.44, p < .01$ ;  $F_2(1, 15) = 23.99, p < .001$ . The display that appeared first received more fixations than the display appearing second. There was also an interaction between all three factors,  $F_1(1, 46) = 4.32, p < .05$ , which was not confirmed by the items analysis,  $F_2(1, 15) = 2.28$ .

Fixation durations were also submitted to a mixed-design ANOVA, revealing an unconfirmed main effect of presentation order,  $F_1(1, 46) = 4.17, p < .05$ ;  $F_2 < 1$ , and a main effect of display type,  $F_1(1, 46) = 87.49, p < .001$ ;  $F_2(1, 15) = 97.15, p < .001$ , but no effect of validity,  $F_1$  and  $F_2 < 1$ . Fixations tended to be longer when the picture appeared first (265 ms) than when it appeared second (248 ms), and fixations were longer on pictures (278 ms) than on sentences

(234 ms). Order of presentation and display type also interacted,  $F_1(1, 46) = 25.89, p < .001$ ;  $F_2(1, 15) = 30.50, p < .001$ , and an analysis of simple main effects revealed an effect of order for sentences,  $F_1(1, 46) = 18.22, p < .001$ ;  $F_2(1, 15) = 5.23, p < .001$ , but not pictures,  $F_1$  and  $F_2 < 1$ . When sentences were read first they attracted shorter fixations (214 ms) than when they were read last (255 ms). Sentences always attracted shorter fixations than pictures, but this advantage was greater when the sentence was read before the picture.

Inspection of sentences is independent of the order of presentation. Inspection times and numbers of fixations are constant, but fixation durations vary. It is perhaps surprising that if durations increase when the number of fixations is unchanged, then the total inspection time does not increase. The data in Table 2 do indicate a slight increase in inspection time however, from 2.02 s (sentence first) to 2.41 s (sentence second), but this effect was not reliable as a function, perhaps, of variations introduced by varying lengths of sentences. The individual fixation durations were, of course, independent of this source of data noise.

These overall analyses indicated a number of effects of the displays upon response and inspection measures, but no effects of validity. This is in contrast to Experiment 1, in which simultaneous presentation of sentence and picture confirmed the established advantage for valid over invalid pairs of stimuli. When participants saw the sentence first in Experiment 2, they read it with shorter fixations than when they saw it second, but the overall amount of attention given to the sentence did not depend upon its order of presentation. Whereas the time taken to read the sentence was constant, the time taken to search the picture depended upon whether they inspected it before or after the sentence. When participants saw the picture first they looked at it for considerably longer (3.92 s) than when they saw after the sentence (1.86 s), reflecting the extensive encoding of a picture in preparation for an unknown question. When the picture was seen after the sentence had been read, then the search to verify the sentence could be focused, with selective inspection of the configuration of the objects highlighted by the sentence. This pattern of results is consistent with the prediction made on the basis of the Henderson et al. (1999) experiments in which a memory task was compared with a search task. In comparing the general encoding necessary for a memory task viewers made more fixations than in the target object search task, similar to the inspection behaviour of viewers seeing the picture first. There was an important difference between the two studies, in addition to the task instructions. The memory task used by Henderson et al. displayed the picture for 15 s, but viewing of the same picture in the search task was self-terminating upon target identification. Consistent presentation conditions were used in the present experiment, with participants pressing a response key to terminate the displays. Henderson et al. also found longer fixation durations when preparing for a recognition task, whereas we did not find a difference in the fixation durations on pictures in the two presentation conditions. If longer fixations were a product of a viewer's encoding processes in the memory task, as suggested by Henderson et al., then we should have seen the same increase here, when pictures were shown first.

## GENERAL DISCUSSION

When extracting information from combinations of sentences and pictures, the viewer's purpose is clearly of importance. A simple verification task (e.g., "the star is below the cross") would be expected to result in a different pattern of eye fixations than a task involving the

understanding of a cartoon (Carroll et al., 1992) or the search for information in a commercial advertisement (Rayner et al., 2001). Perhaps surprisingly then, common patterns of inspection do emerge when text and graphics are available together, with fixation durations tending to be shorter on the textual components than on the graphics, and the order of search involving few fixations on the graphics followed by more extensive inspection of the text and then further examination of the scene.

In the present experiments the task was to verify the truth of a sentence as a description of the events shown in a photograph of a natural scene. The sentence was shown at the same time and directly below the picture in Experiment 1, and in Experiment 2 the sentence either preceded or followed the picture. Fixation durations, taken as an index of processing difficulty, were shorter on sentences than on pictures in both experiments. The number of fixations made upon the two components of the display varied according to their order of presentation and can be interpreted as the product of differences between general encoding and directed search. When the sentence appeared first, or when it was available at the same time as the picture, then there were more fixations on the sentence than on the picture. This is an effect of variations in the inspection of the picture rather than variations in reading patterns. The number of fixations on the sentences remained relatively constant, regardless of presentation type (see Tables 1 and 2), but the durations of those fixations did vary. If the viewers first read the sentence and could focus their search of the picture, then fewer fixations were made on the picture than in those presentations where it was presented first and had to be encoded as a complete scene. Table 2 shows that when the picture appeared last it attracted approximately half the number of fixations than when it appeared first. This is not the pattern reported by Carroll et al. (1992), where the number of fixations on the picture remained constant over presentation order, presumably because the task of understanding the cartoon invariably demanded detailed inspection of the same objects in the line drawing. It is consistent with a prediction made on the basis of the Henderson et al. (1999) experiments, however, with more fixations when viewers were attempting to encode the whole scene rather than search it for a target object.

When the picture was presented prior to the sentence, then the viewer had the task of encoding as many features and relationships as possible, in anticipation of an unknown question. The task was to understand and encode the events shown in the picture, and this resulted in longer inspection of the picture (3.92 s) than when the picture appeared after the sentence (1.86 s). When the picture appeared after the sentence, inspection could focus upon the most salient features, and this was also indicated by an increased number of fixations on the picture when it appeared first rather than second. In contrast, the inspection time and number of fixations on the sentence remained constant regardless of the order of presentation. Understanding the sentence presented the same difficulty to readers whether it appeared first or last, according to these measures, but fixation durations suggest that there was an increased difficulty when it appeared last. Fixations on sentences were shorter when they appeared before the picture than when they appeared after it. Encoding the relationships between referents previously seen in the picture results in an increase in fixation durations. This is an effect upon comprehension of the same sentence read before or after the picture that is to be interrogated. Prior to seeing the picture, the sentence requires the construction of an abstract propositional model, and relatively short fixations are observed. When it is read after seeing a picture, however, the task is one of relating the same referents

to the objects in an encoded scene. The extra processing time reflects the additional process of comparing referents across media.

Comparing the patterns of results from concurrent displays (Experiment 1) with those from sequential displays (Experiment 2) provides a closer match when looking at those trials when sentences preceded the pictures. There was no difference between the total inspection time on sentences and pictures in either experiment. With concurrent displays sentences and pictures received similar amounts of attention as when, in Experiment 2, the sentences preceded the pictures. Comparing the number of fixations made in the two experiments also provides a similar pattern, with sentences receiving more fixations than pictures. With simultaneous displays of sentences and pictures, there were similar numbers of fixations as when the sentence preceded the pictures. When the sentence could be read first, as it must be with successive presentations of sentence–picture and as is possible with concurrent displays, then sentences do not receive more attention overall, but they do receive more fixations than the pictures. This apparent paradox is resolved by differences in fixation durations between sentences and pictures. With concurrent displays the sentences received shorter fixations than the pictures, and this was also the case when sentences preceded pictures. When the sentence can be read first then inspection of the picture can be specific to the foregrounded relationship between the depicted objects, resulting in fewer fixations on the picture than on the sentence. The recognition of words in sentences requires less processing time and shorter fixations than does the recognition of objects in pictures (see also Carroll et al., 1992; Rayner et al., 2001).

When sentences were presented first (Experiment 2), performance resembled that in Experiment 1, in which the picture and the sentence appeared together. The exception to this generalization is in the absence of a falsification effect with successive displays. The *true* trials with concurrent displays gained fewer fixations and a reduced inspection time than the *false* trials, in agreement with past studies of sentence–picture verification tasks. There was no effect upon fixation duration, implying that the effect is located in post-encoding decision processes rather than in a perceptual stage of processing. Why was there no falsification effect with successive sentence–picture displays?

In sentence verification tasks involving simple displays such as geometrical shapes accompanied by active declarative statements, the established pattern of results is that *true* statements gain faster responses than *false* statements (and if the statement contains a negative, it will be slower than when it does not). This falsification effect was observed in Experiment 1, using concurrent pictures and sentences. The keyboard decision time is the measure most comparable to that taken in the studies by Clark and Chase (1972), Gough (1965), Just and Carpenter (1971), and others, and the advantage for *true* sentences was observed here. This advantage was also seen in the total inspection time (accumulated duration of all fixations) and in the number of fixations made. When the sentence and picture were displayed one after the other, in Experiment 2, then the falsification effect was not apparent. There was only a marginal difference in the number of fixations made on disconfirming pictures when they followed the sentence, but this was supported neither by the keyboard response time nor by the accumulated inspection time. Experiment 2 found that when the sentence appears after offset of the picture, and when the picture appears after offset of the sentence, there was no difference in the responses to *true* and *false* statements. This is contrary to an established literature, and contrary to the results from Experiment 1 in which the picture and sentence appeared concurrently.

The advantage of *true* statements over *false* statements is one of the principal effects that models of sentence verification were designed to explain (Carpenter & Just, 1975; Clark & Chase, 1972). The pattern was not seen in Experiment 2 here, and it was not seen in all conditions of the experiments reported by Goolkasian (1996, 2000) and Feeney et al. (2000). In Goolkasian's experiments the effect was only present when the task required a verbatim comparison between picture and sentence, and in the Feeney et al. experiment the effect required the matching of referents in the sentence, and the graph was a better predictor of verification time than of the accuracy of the sentence. These experiments question the assumption that performance of the sentence verification task requires the construction of comparable abstract propositional forms from the sentence and the picture. In contrast to Clark and Chase's supposition that the sentence and the picture are represented in the same propositional form, Larkin and Simon (1987) have argued that although they may contain the same information, the processing operations required to extract the information will not necessarily be equivalent. Pictures and diagrams have advantages over textual descriptions—for example, in the perception of relationships between elements of a diagram and in the efficiency of the search for specific elements. These recognition processes result in performance gains for graphical representations over textual forms that are propositionally equivalent. Evidence in support of this model was provided by Goolkasian (2000), who found that slower decisions were made when the information was presented as text rather than simple pictures. When the stimulus onset asynchrony (SOA) between a picture and the test sentence was varied, the size of the pictorial advantage decreased as the SOA decreased, suggesting that the format influenced extraction of information. In the presentations of pictures of natural scenes and short declarative sentences in the present study, we found longer fixation durations on pictures than on sentences, a pattern consistent with results from Carroll et al. (1992) and Rayner et al. (2001). The easier recognition of relationships in a picture, proposed by Larkin and Simon (1987), is not reflected in fixation durations. We can conclude that the richer representations of information in pictures require extended encoding durations relative to the encoding of information from text.

When inspecting combinations of pictures and text, processing is easier when the text is read first. This generalization is supported by all four measures taken in Experiment 2, where the two components were presented one after the other. In addition, when the picture and text are presented together, participants tend to look at the sentence first. Longer fixations are associated with more extensive processing, and on this basis it can be concluded that the rich encoding of objects shown in pictures require more processing than the words in sentences. When searching Figure 1 for the cyclist mentioned in the accompanying sentence, we find that the car is actually overtaking a horse, but what we encode can be a much richer representation than suggested by this single word—the horse is walking along the side of a rural road, it is a bay, it is heading away from us, it has a rider who is wearing a helmet, and so on. This encoding of this detail requires the extended fixation durations seen here and in previous studies. Presenting the sentence prior to the picture allowed faster processing of the sentence than when it appeared after the picture, and this was seen in the overall inspection time as well as in the durations of fixations. When participants read the sentence first they were able to make a selective search of the picture to identify the referents and their relationships. A surprising feature of the present results is the failure to confirm the traditional result from sentence verification tasks, in which *false* instances gained slower responses than *true* instances. This

pattern was only seen here when sentence and picture appeared together. When the same stimuli were presented separately, there was no difference between *true* and *false* trials. This may be due to the complexity of the pictorial stimuli used here, in contrast with the simple geometric displays used previously. The pictures of natural scenes perhaps precluded the generation of an abstract propositional model of the relationships between referents because there were a large number of possible relationships to be encoded. Simple displays may allow a propositional analysis and may have encouraged the development of models of verification processes that are restricted in their scope.

## REFERENCES

- Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, *82*, 45–73.
- 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.
- Chapman, P. R., & Underwood, G. (1998). Visual search of driving situations: Danger and experience. *Perception*, *27*, 965–976.
- Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472–517.
- Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240–247.
- Crundall, D. E., & Underwood, G. (1998). Effects of experience and processing demands on visual information acquisition in drivers. *Ergonomics*, *41*, 448–458.
- Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*, 641–655.
- Feeney, A., Holo, A. K. W., Liversedge, S. P., Findlay, J. M., & Metcalf, R. (2000). How people extract information from graphs: Evidence from a sentence–graph verification paradigm. In M. Anderson, P. Cheng, & V. Haarslev (Eds.), *Theory and application of diagrams: First international conference, Diagrams 2000* (pp.149–161). Berlin: Springer-Verlag.
- Goolkasian, P. (1996). Picture–word differences in a sentence verification task. *Memory & Cognition*, *24*, 584–594.
- Goolkasian, P. (2000). Pictures, words, and sounds: From which format are we best able to reason? *Journal of General Psychology*, *127*, 439–459.
- Gough, P. B. (1965). Grammatical transformations and speed of understanding. *Journal of Verbal Learning and Verbal Behavior*, *5*, 107–111.
- Hegarty, M. (1992a). The mechanics of comprehension and comprehension of mechanics. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 428–443). New York: Springer-Verlag.
- Hegarty, M. (1992b). Mental animation: Inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1084–1102.
- 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.
- Kennedy, A., Radach, R., Heller, D., & Pynte, J. (2000). *Reading as a perceptual process*. Oxford, UK: Elsevier.
- Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with qualification. *Journal of Verbal Learning and Verbal Behavior*, *10*, 244–253.
- Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth a thousand words. *Cognitive Science*, *11*, 65–99.
- 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.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372–422.

- Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14*, 191–201.
- 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.
- Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science, 8*, 374–378.
- Underwood, G. (1998). *Eye guidance in reading and scene perception*. Oxford, UK: Elsevier.
- Underwood, G., Chapman, P., Bowden, K., & Crundall, D. (2002). Visual search while driving: Skill and awareness during inspection of the scene. *Transportation Research (F): Traffic Psychology and Behaviour, 5*, 87–97.
- Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology, 11*, 92–107.
- Wason, P. C., & Jones, S. (1963). Negatives: Denotation and connotation. *British Journal of Psychology, 54*, 92–107.

*Original manuscript received 6 June 2002*

*Accepted revision received 9 December 2002*

*PreView proof published online 16 June 2003*