Available online at www.sciencedirect.com
SCIENCE DIRECT®
COGNITION
ELSEVIER
Cognition 93 (2004) B79–B87
www.elsevier.com/locate/COGNIT

Brief article

# Language-mediated eye movements in the absence of a visual world: the 'blank screen paradigm'

Gerry T.M. Altmann*

*Department of Psychology, University of York, Heslington, York YO10 5DD, UK*

Received 5 September 2003; revised 12 January 2004; accepted 9 February 2004

**Abstract**

The 'visual world paradigm' typically involves presenting participants with a visual scene and recording eye movements as they either hear an instruction to manipulate objects in the scene or as they listen to a description of what may happen to those objects. In this study, participants heard each target sentence only after the corresponding visual scene had been displayed *and then removed*. For a scene depicting a man, a woman, a cake, and a newspaper, the eyes were subsequently directed, during '*eat*' in '*the man will eat the cake*', towards where the cake had previously been located even though the screen had been blank for over 2 s. The rapidity of these movements mirrored the anticipatory eye movements observed in previous studies [Cognition 73 (1999) 247; J. Mem. Lang. 49 (2003) 133]. Thus, anticipatory eye movements are not dependent on a concurrent visual scene, but are dependent on a mental record of the scene that is independent of whether the visual scene is still present.


*Keywords:* Sentence comprehension; Eye movements; Visual scene interpretation

## 1. Introduction

Eye movements are closely time-locked to the occurrence of spoken expressions that enable the identification of items depicted by a concurrent visual scene (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Altmann and Kamide (1999), for example, presented participants with a scene depicting a boy, a cake, and various other items, and monitored participants' eye movements as, concurrently, they heard the sentence '*the boy will eat the cake*'. They observed more saccades towards the cake (the one edible

* Tel.: +44-1904-434362; fax: +44-1904-433181.
*E-mail address:* g.altmann@psych.york.ac.uk (G.T.M. Altmann).

0022-2860/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2004.02.005

item in the scene) during the acoustic lifetime of '*eat*' than when this verb was replaced by another whose selectional restrictions were more general (e.g. '*move*'). Such *anticipatory* eye movements suggest that the processing system can rapidly anticipate properties of what will be referred to post-verbally (cf. Kamide, Altmann, & Haywood, 2003a). However, although language is often used in visual contexts containing the items to which that language refers, it is commonly used also when those items are absent.

Recent studies suggest that the spatial properties of absent stimuli can influence eye movements when those stimuli are imagined or when information associated with those stimuli is recalled (e.g. Brandt & Stark, 1997; Laeng & Teodorecu, 2002; Richardson & Spivey, 2000; Spivey & Geng, 2001; see Spivey, Richardson, & Fitneva, 2004, for review). For example, Richardson and Spivey (2000) have shown that the eyes will, under certain circumstances, move to where information had previously been introduced; a different face appeared at different positions within a grid, and each presented some fact (e.g. a fact about Shakespeare) before disappearing. Subsequently, when information about one of these facts was given to the participants, their eyes moved back to where in the array the face associated with that fact had been located.

The present study, unlike these previous ones, seeks to determine whether, *and with what time-course*, the eyes move to where specific items had been located when, subsequently, those items are directly referred to. The study is similar to that of Altmann and Kamide (1999), with one critical change: instead of presenting the visual scene and the target sentence concurrently, the visual scene was presented to participants first, but then taken away *before* the target sentence was played. If anticipatory eye movements are contingent on mental representations of the items in a scene, independently of their physical presence, equivalent movements might be observed if the visual scene is removed before participants hear the target sentence.

## 2. Method

### 2.1. Subjects

Thirty native speakers of English from the University of York student community took part in this study.

### 2.2. Stimuli

Twenty experimental pictures (see Fig. 1) were each paired with two sentential conditions corresponding to (1) & (2) below.

(1) The man will eat the cake
(2) The woman will read the newspaper

Each scene portrayed two protagonists and two items—one satisfying the selection restrictions of one verb, and another satisfying the selection restrictions of the other. Each scene was followed by either one or other of the pair of sentences. By comparing looks

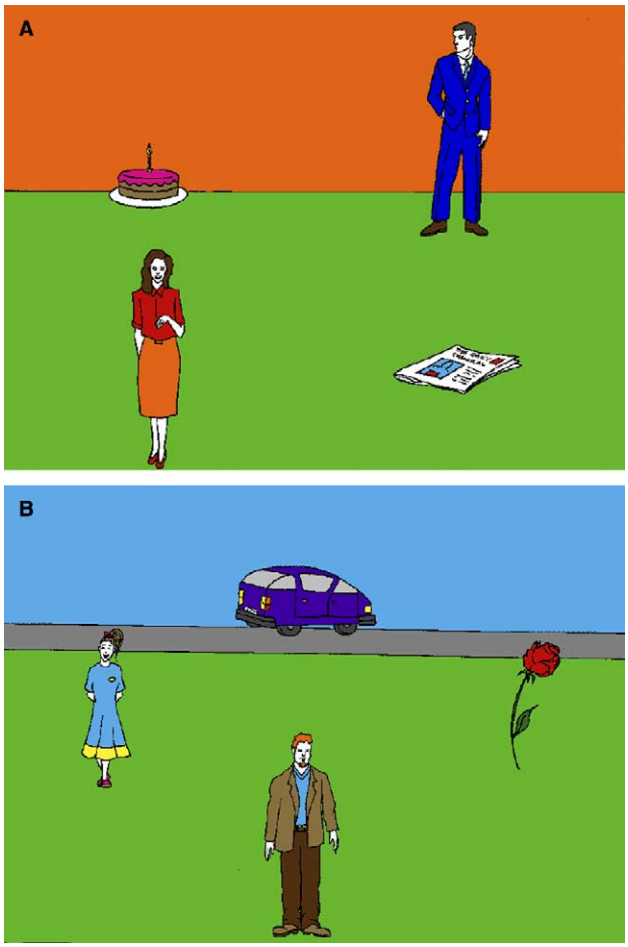


Fig. 1. Example scenes. The accompanying sentences were: (A) '*The man will eat the cake/the woman will read the newspaper*'; (B) '*The girl will smell the flower/the man will drive the car*'.

towards the cake's prior location with looks towards the newspaper's prior location after '*the man will eat the cake*', and looks towards the newspaper's prior location with looks towards the cake's prior location after '*the woman will read the newspaper*', each item served as its own control (looks towards the 'inappropriate' item, as determined by the verb, are the baseline against which to compare looks towards the 'appropriate' item). The design also ensures that the potential effects of any differences in visual salience amongst the different objects are eliminated.

For half the scenes, the four target items occupied distinct quadrants of the display (cf. Fig. 1A). For the other half, they were in a 'diamond' configuration (cf. Fig. 1B). A further 20 sentence/scene pairs were added as fillers. These employed similar scenes but included a range of other sentence types, with ditransitive verbs and verbs that did not determine, given the scene, which item would be referred to next.

Table 1
Extents and durations of regions of analysis

| Label | Extent | Duration (ms) |
|---|---|---|
| the N1 | onset of '*the*' to offset of '*man*' or '*woman*' | 573 |
| will | offset of '*man*' or '*woman*' to onset of '*eat*' or '*read*' | 608 |
| VERB | onset of '*eat*' or '*read*' to onset of '*the*' | 826 |
| | (verb onset to verb offset | 441) |
| | (verb offset to determiner onset | 385) |
| the N2 | onset of '*the*' to offset of '*cake*' or '*newspaper*' | 536 |

Two lists of stimuli, in fixed-random order, were created containing each of the 20 experimental pictures but just one version of each sentence pair. The sentences were recorded by a male native speaker of British English (GTMA), and sampled at 44.1 kHz. The prosody of each utterance was normal, but slightly exaggerated to ensure 'clear and careful' speech (hence the slight pause after the verb; see Table 1). The scenes were presented on a 17″ viewing monitor at a resolution of $640 \times 480$ pixels.

### 2.3. Procedure

Participants were seated with their eyes between 20″ and 25″ from the display. They wore an SMI EyeLink head-mounted eye-tracker, sampling at 250 Hz. Participants were given the following instructions: *In this experiment you will be asked to look at some pictures on a computer screen. Following each picture will be a short sentence spoken over the loudspeakers. The picture might show for example two people and various objects. The sentence might be like "the boy will beat the drum". This experiment is interested in what happens when a person hears a sentence describe something that might happen in a picture that was presented earlier.* Participants were not asked to perform any explicit task, thereby avoiding mention of meta-linguistic judgments, memorization, or speed of processing, each of which might induce unnatural processing strategies. Altmann and Kamide (1999, Expt. 1) asked participants to judge whether each sentence could in principle apply to the concurrent scene, but a number of other studies have employed the current 'no-task task' and replicated the Altmann and Kamide (1999) result (e.g. Altmann & Kamide, 1999, Expt. 2, 2004; Kamide et al., 2003a; Kamide, Scheepers, & Altmann, 2003b). Each visual stimulus remained onscreen for 5 s and was then replaced by a plain white screen. One second later (cf. Spivey & Geng, 2001), with the screen still blank (and without any instruction to fixate any particular location), the target sentence was presented. The trial was automatically terminated 5 s after the onset of the target sentence.

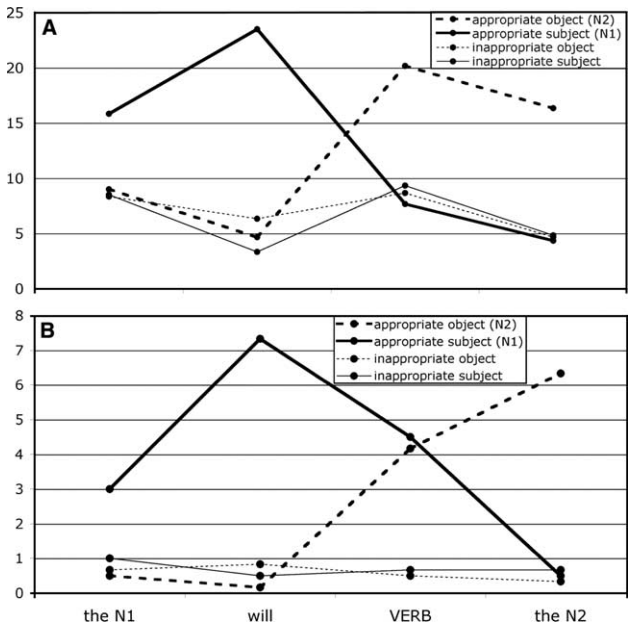


Fig. 2. Percentages of trials with saccadic eye movements directed towards the prior locations of the target items. Panel A depicts the quadrant analyses, and Panel B the pixel analyses.

## 3. Results

Looks will be reported towards where the appropriate subject had been (the man for '*the man will eat the cake*' and the woman for '*the woman will read the newspaper*') or to where the appropriate object had been (the cake or the newspaper, depending on the target sentence).

Two distinct analyses of the data are reported. In the first, a saccadic movement was counted if the resulting fixation fell within the quadrant that had contained the target item. In the second, saccades were counted only if the resulting fixation landed on the actual pixels previously occupied by that target item. For the sake of exposition, 'looks towards the appropriate item' will mean looks towards where that target item *had been located*. The graphs in Fig. 2 show the percentage of trials with at least one look to each target item, synchronized against the speech input. This is the same measure used in the statistical analyses reported by Altmann and Kamide (1999); whereas they were interested in just the verb region, we are here interested in the two noun phrase regions also (see Altmann & Kamide, 2004, for discussion of the equivalence across different measures, and alternative ways of synchronizing and graphing the data). The extent of each temporal region (e.g. the verb region, defined here as between verb onset and onset of the following determiner) was determined on a trial-by-trial basis—see Table 1 for average durations. Because we report percentages of saccades, we eliminate any trials on which the participant is already fixating the critical location at the onset of the temporal region of interest (but see

footnote1). *F*-tests were performed on the arcsine-transformed data. However, and given that the number of trials with saccades towards the target region was very low in some cases, the non-parametric Wilcoxon test was also applied; unless stated differently, all statistical patterns were confirmed by the Wilcoxon test ($\alpha = 0.05$).

### 3.1. Quadrant analyses

There were significantly more trials during '*the N1*' with looks towards the appropriate subject than towards the inappropriate subject: $F1(1, 29) = 6.5$, $P < 0.02$; $F2(1, 19) = 24.8$, $P < 0.0001$. There were no more looks towards the appropriate object than towards the inappropriate object (both $F < 1$). This same pattern persisted during '*will*': $F1(1, 29) = 42.2$, $P < 0.0001$; $F2(1, 19) = 92.0$, $P < 0.0001$, and for looks towards the appropriate/inappropriate objects, both $F < 1$. During the verb, there were more looks towards the appropriate object than towards the inappropriate object: $F1(1, 29) = 11.8$, $P < 0.002$; $F2(1, 19) = 17.7$, $P < 0.001$. There were no more looks during this region towards the appropriate subject than towards the inappropriate subject (both $F < 1$). This same pattern persisted during '*the N2*': $F1(1, 29) = 19.1$, $P < 0.0002$; $F2(1, 19) = 30.4$, $P < 0.0001$, and for looks towards the appropriate/inappropriate subject, both $F < 1$.

### 3.2. Pixel analyses

There were no differences during '*the N1*' between looks towards the appropriate and inappropriate subjects ($F1(1, 29) = 2.2$, $P > 0.1$; $F2(1, 19) = 3.9$, $P < 0.07$) or towards the appropriate and inappropriate objects (both $F < 1$). Wilcoxon tests suggested a marginal effect for looks to the subjects (significant by items only). During '*will*' there were significantly more looks towards the appropriate subject: $F1(1, 29) = 19.6$, $P < 0.0002$; $F2(1, 19) = 45.0$, $P < 0.0001$. There were marginally more looks towards the inappropriate object than towards the appropriate object ($F1(1, 29) = 2.8$, $P > 0.1$; $F2(1, 19) = 4.7$, $P < 0.05$), but Wilcoxon tests were not significant. During the verb, there were still more looks directed towards the appropriate subject than towards the inappropriate subject ($F1(1, 29) = 20.5$, $P < 0.0001$; $F2(1, 19) = 23.7$, $P < 0.0002$), but there were also more looks towards the appropriate object than towards the inappropriate object: $F1(1, 29) = 17.3$, $P < 0.0003$; $F2(1, 19) = 25.2$, $P < 0.0001$. This latter pattern persisted during '*the N2*' ($F1(1, 29) = 43.0$, $P < 0.0001$; $F2(1, 19) = 35.5$, $P < 0.0001$), but there were no more looks towards the appropriate subject than towards the inappropriate subject (both $F < 1$).

The graphs in Fig. 2 show data averaged across distinct trials, and looks may have been directed to the appropriate subject's prior locations on some trials, and towards the appropriate objects' prior locations on others. They do not necessarily reflect, as the target sentence unfolded, looks *from* one *to* the other. However, a contingent analysis of just those trials in which participants did initiate a saccade towards the appropriate subject between sentence onset and verb onset (39% of trials) revealed that on 29% of these trials looks were subsequently directed towards the appropriate object during the verb (up until the onset of the post-verbal article), and that on a further 20% of these trials, looks were

directed towards the appropriate object during the post-verbal referring expression. Thus, on almost 50% of occasions when the eyes moved towards the prior location of the appropriate subject, the eyes subsequently moved, after verb onset, towards the prior location of the appropriate object.[1]

## 4. Discussion

As participants heard particular items being mentioned, so they looked towards where those items had been located. Moreover, looks towards the previous location of whatever would be referred to in grammatical object position were launched during the verb region itself, and this was observed in both the quadrant and pixel analyses.[2,3] The data thus replicate the verb-based patterns reported by Altmann and Kamide (1999). Because that study employed a different design (it manipulated verb selection restrictions), and different stimuli (different and more items, arranged non-geometrically), the two sets of data are not directly comparable. Nonetheless, both sets of data indicated that looks towards the appropriate locations relative to inappropriate locations were initiated during the verb (on 20% of trials between verb onset and determiner onset in the quadrant analyses reported here, and on approximately 28% of trials in Altmann and Kamide (1999), Expt. 2).

Why did the eyes move to a particular location when there was nothing there? One possibility is based on the idea that very little information about one part of a visual scene is maintained internally when the eyes move to another part. Richardson and Spivey (2000) proposed, following O'Regan (1992), that the visual system instead uses the scene itself as an external memory, using oculomotor coordinates (defined relative to the configuration of cues within the scene) as pointers towards this external memory (cf. Ballard, Hayhoe, Pook, & Rao, 1997). The activation of these pointers causes the eyes to move to the corresponding coordinate from where information about the contents of that part of the scene can be retrieved. The effects we observed above can be explained in terms of a system in which linguistic expressions drive eye movements

[1] The figure of 39% excludes occasions when the eyes were already in the appropriate quadrant at sentence onset. An analysis which included such fixations revealed that on 63% of trials, participants had fixated the location of the appropriate subject between sentence onset and verb onset. On 42% of these trials, participants moved towards the location of the appropriate object between verb onset and sentence offset. On the majority of the remaining trials, participants continued to fixate the location of the subject.

[2] Analyses between verb onset and verb offset (as opposed to determiner onset) suggested more looks towards where the appropriate object had been than towards where the inappropriate object had been—quadrant analyses: $F1(1,29) = 4.3$, $P < 0.05$; $F2(1,19) = 4.8$, $P < 0.05$; pixel analyses: $F1(1,29) = 5.6$, $P < 0.03$; $F2(1,19) = 7.8$, $P < 0.02$. However, only the pixel analyses were significant on the Wilcoxon tests.

[3] A replication of this study was subsequently undertaken with Adele Pacini, with half the participants being given approximately 0.4 g/kg alcohol prior to taking part in the experiment. This gives a peak blood alcohol concentration of 0.07 (0.05 is the legal limit in most European countries; 0.08 is the legal limit in the UK and Italy, but is considered over the legal limit in most US states). The pattern of results was identical to that reported here, although there was a lag of approximately 100 ms in the onset of the eye movements of participants in the 'alcohol' condition relative to those in the 'no alcohol' condition. The effect is thus robust even during mild intoxication.

towards particular locations on the premise that further information about whatever is in that location might be retrieved once the eyes reach their destination. One might suppose, however, that when the screen is entirely blank, the processing system would be aware that there was nothing relevant in the visual field. In which case, why move the eyes? And why with such a rapid time-course? Richardson and Spivey (2000) concluded that "saccades to blank regions of space [are] a consequence of a perceptual-motor system that relies on rapid access to an external store of information, and doesn't always know when that external store has changed." (p. 292). That the system does not always know when that external store has changed may be true when only a part of the scene changes (cf. Spivey & Geng, 2001), but is less likely when the entire scene is removed. More likely is the idea (also considered by Richardson & Spivey, 2000) that the spatial pointers are a component of the episodic trace associated with each item—activating that trace necessarily activates the (experiential) component encoding the location of that item, and it is this component that automatically drives the eyes towards that location.

In conclusion: Eye movements that are triggered during linguistic expressions are not contingent on an item being co-present with that expression. Thus, even when the visual scene *is* concurrent with a linguistic expression that refers to an item within that scene, information about where to move the eyes in order to fixate that item may be based not on the actual location of that item within the scene, but on the location of that item as represented within a mental representation of the scene. On the basis of data reported elsewhere (e.g. Altmann & Kamide, 2004), we believe that this representation is an *interpreted* representation that can dynamically change as a function of the linguistic input. How much information about the objects in a visual scene is represented apart from the scene (within such representations), and to what level of detail, is the subject of considerable debate (see Henderson & Ferreira, 2004, for review)—language-mediated eye movements in the absence of (previously seen) visual scenes may usefully inform this debate.

Perhaps the most surprising aspect of the data, and one that has not been revealed in previous studies, is their time-course: Eye movements were launched at the earliest moments, both during the referring expressions themselves (e.g. '*the man*/*the cake*') and during the verb ('*eat*'). Thus, the available evidence suggests that the time-course of the mapping between language and the visual world is largely the same, whether that visual world is just a memory, or is still 'out there'.

## Acknowledgements

This research was supported by the Medical Research Council (grants G9628472N and G0000224), the Economic and Social Research Council (R000222798), and the Royal Society. The author thanks Yuki Kamide for programming the eye-tracker, Andrew Thomas for creating the original visual scenes, and Ruth Ancliff for running the present study. A partial report of these data appears in Altmann and Kamide (2004).

## References


Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Altmann, G. T. M., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. M. Henderson, & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioural and Brain Sciences*, *20*(4), 723–767.

Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, *9*, 27–38.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107.

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson, & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003a). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*, 133–159.

Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003b). Integration of syntactic and semantic information in predictive processing: A cross-linguistic study in German and English. *Journal of Psycholinguistic Research.*, *32*(1), 37–55.

Laeng, B., & Teodorecu, D. (2002). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cognitive Science*, *26*, 207–231.

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

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood squares: Looking at things that aren't there anymore. *Cognition*, *76*, 269–295.

Spivey, M. J., & Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological Research*, *65*, 235–241.

Spivey, M. J., Richardson, D. C., & Fitneva, S. A. (2004). Thinking outside the brain: Spatial indices to visual and linguistic information. In J. M. Henderson, & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.