Running head: EYE MOVEMENTS AND FREE RECALL

The Effect of Horizontal Eye Movements on Free Recall: A Preregistered Adversarial Collaboration

Dora Matzke[1], Sander Nieuwenhuis[2], Hedderik van Rijn[3], Heleen A. Slagter[1], Maurits W. van der Molen[1], and Eric-Jan Wagenmakers[1]

[1] University of Amsterdam

[2] Leiden University

[3] University of Groningen

Correspondence concerning this article should be addressed to:

Dora Matzke

University of Amsterdam, Department of Psychology

Weesperplein 4

1018 XA, Amsterdam, The Netherlands

Phone: +31205258862

E-mail to d.matzke@uva.nl.

Word count: 8,380

Abstract

A growing body of research suggests that horizontal saccadic eye movements facilitate the retrieval of episodic memories in free recall and recognition memory tasks. Nevertheless, a minority of studies have failed to replicate this effect. The present paper attempts to resolve the inconsistent results by introducing a novel variant of proponent-skeptic collaboration. The proposed approach combines the features of adversarial collaboration and purely confirmatory preregistered research. Prior to data collection, the adversaries reached consensus on an optimal research design, formulated their expectations, and agreed to submit the findings to an academic journal regardless of the outcome. To increase transparency and secure the purely confirmatory nature of the investigation, the two parties set up a publicly available adversarial collaboration agreement that detailed the proposed design and all foreseeable aspects of the data analysis. As anticipated by the skeptics, a series of Bayesian hypothesis tests indicated that horizontal eye movements did not improve free recall performance. The skeptics suggest that the non-replication may partly reflect the use of suboptimal and questionable research practices in earlier eye movement studies. The proponents counter this suggestion and use a p-curve analysis to argue that the effect of horizontal eye movements on explicit memory does not merely reflect selective reporting.

Keywords: adversarial collaboration, Bayes factor, horizontal eye movements, preregistration, replication

The Effect of Horizontal Eye Movements on Free Recall: A Preregistered Adversarial Collaboration

The authors declare that they have reported all measures, experimental conditions, data exclusions, and disclosed how they determined their sample size and the corresponding data collection stopping rule.

Introduction

Do horizontal saccades make it easier for people to retrieve events from memory? Past research seems to suggest that they do. A growing number of investigations report that only 30 seconds of horizontal saccadic eye movements can improve memory retrieval and boost performance in both recall and recognition tasks. A number of studies have, however, failed to replicate the seemingly well-established effect of horizontal eye movements on free recall performance.

Motivated by the inconsistent results, here we describe a purely confirmatory proponent-skeptic collaboration that focuses on the association between horizontal eye movements and episodic memory. Proponent-skeptic collaboration has been repeatedly advocated as a constructive method of scientific conflict resolution (Hofstee, 1984; Kahneman, 2003; Latham, Erez, & Locke, 1988; Mellers, Hertwig, & Kahneman, 2001). Moreover, we believe that adversarial collaborations—especially when coupled with the preregistration of the statistical analyses—may remedy a number of factors that contributed to the recent crisis of confidence in psychological research and may increase the transparency of scientific communication (see also Koole & Lakens, 2012; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011).

Preregistered Adversarial Collaboration: A Confirmatory Proponent-Skeptic Investigation

Adversarial collaboration is a cooperative research effort that is undertaken by two (groups of) investigators who hold different views on a particular empirical question. The term adversarial collaboration was coined by Kahneman (2003, see also Latham et al., 1988), who—unsatisfied with the inefficiency of traditional scientific disputes—urged researchers to engage in a "good-faith effort to conduct debates by carrying out joint research" (p. 729). The goal of an adversarial collaboration is to reach consensus on an

experimental design and the corresponding testable hypotheses. To facilitate the interpretation of the results, the adversaries are required to formulate and document their expectations about the outcome of the study prior to data collection. Adversarial collaborations are often carried out under the guidance of a third-party researcher, the arbiter, who oversees the collaboration and acts as an impartial referee in case of disagreements (see also Mellers et al., 2001; Nier & Campbell, 2012). Although adversarial collaboration does not necessarily result in the complete resolution of the disagreement, it often leads to new testable hypotheses and is therefore likely to advance the debate.

Although the past two decades have witnessed a number of successful adversarial collaborations in various disciplines (e.g., Bateman, Kahneman, Munro, Starmer, & Sugden, 2005; Cadsby, Croson, Marks, & Maynes, 2008; Gilovich, Medvec, & Kahneman, 1998; Mellers et al., 2001; Schlitz, Wiseman, Watt, & Radin, 2006; Wiseman & Schlitz, 1997, 1998), this form of conflict resolution is unfortunately still the exception rather than the rule. The lack of adversarial collaboration is especially unfortunate in light of the recent "crisis of confidence" (Pashler & Wagenmakers, 2012, p. 528) in psychological research. The crisis is fueled by concerns about the replicability of key results (e.g., Hunter, 2001) and the widespread use of questionable research practices such as the selective reporting of significant results (e.g., Simmons, Nelson, & Simonsohn, 2011). The controversy has drawn widespread public attention and triggered a broad range of responses. At one end of the spectrum, failures to replicate key studies in the psychological literature (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012; Shanks et al., 2013) have prompted hostility and finger-pointing between research groups. At the other end of the spectrum, the dispute has given rise to valuable attempts to identify and remedy the factors that contributed to the development of the crisis. Although the proposed recommendations vary considerably in focus, they all emphasize the importance of increasing the transparency of scientific communication (Ioannidis, 2005; Koole &

Lakens, 2012; Pashler & Harris, 2012; Wagenmakers et al., 2011; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

Transparency should not only be a concern once the data have been collected; it has been suggested that researchers should commit themselves to the methods of data analysis prior to data collection (e.g., Wagenmakers et al., 2012; de Groot, 1956/2014; de Groot, 1969). Failure to do so may lure researchers into tailoring the analyses to patterns in the observed data in order to find statistically significant results (John, Loewenstein, & Prelec, 2012; Simmons et al., 2011). Fishing for significant results, however, invalidates the interpretation of Type I and Type II error rates and may lead to distorted statistical conclusions. In fact, Wagenmakers et al. (2012) argued that the widespread confusion between exploratory and confirmatory research is the main "fairy-tale" factor in contemporary psychology. Wagenmakers et al. have therefore urged researchers to preregister their studies and publicly disclose prior to data collection which dependent variables they intend to measure and which statistical analyses they intend to conduct (see also Bakker, van Dijk, & Wicherts, 2012; Chambers, 2013; Chambers, Munafo, & et al., 2013; de Groot, 1956/2014; Goldacre, 2009; Ioannidis, 2005; Koole & Lakens, 2012; Nosek, Spies, & Motyl, 2012; Wagenmakers et al., 2011; Wolfe, 2013). The preregistration of experiments has been substantially simplified by the development of web-based research archives and data repositories such as the Open Science Framework (OSF; `http://osf.io`).

Here we advocate a hybrid variant of scientific conflict resolution that combines the features of adversarial collaboration (Kahneman, 2003) and preregistered confirmatory research (Wagenmakers et al., 2012). The proposed approach may not only assists the constructive resolution of scientific debates, but may also remedy a number of factors that contributed to the present crisis in psychology. We propose the following guidelines for preregistered proponent-skeptic collaborations (see also Mellers et al., 2001, and Hofstee,

1984, for suggestions on adversarial collaborations). First, the adversaries reach consensus on an optimal research design. This precaution eliminates the possibility of later disputes regarding the execution of the study. Second, the two parties formulate their hypotheses and expectations in advance. This precaution decreases the probability of the investigators falling prey to various cognitive biases, such as hindsight bias (i.e., judging an event as more predictable after it has occurred; Roese & Vohs, 2012) and confirmation bias (i.e., favoring information that confirms one's own hypotheses; Nickerson, 1998). Third, the adversaries agree to write a joint article and submit it to an academic journal regardless of the outcome of the study. This precaution may in the long term counteract publication bias and the file drawer problem (Rosenthal, 1979; Greenwald, 1975). Lastly, as the novel but crucial ingredient, the two parties set up an adversarial collaboration agreement. The agreement describes the proposed research design and all foreseeable aspects of the pre-processing and analysis of the data. This precaution secures the purely confirmatory nature of the investigation and increases the transparency of scientific communication.

The remainder of the article describes a joint investigation that focused on the effects of horizontal eye movements on episodic memory. We will first introduce the research area, motivate the reasons for the preregistered adversarial collaboration, and describe the proposed experimental design and the corresponding statistical analyses. We will then describe the methods of the study in more detail and present the results of the investigation. Lastly, the adversaries will present their own perspective on the results as well as on the process of the joint work.

Horizontal Eye Movements and Episodic Memory

Background and Motivation

Past research suggests that horizontal saccadic eye movements assist the consolidation and retrieval of memories. For instance, bilateral eye movements have been

reported to decrease the severity of memory symptoms in eye-movement desensitization and reprocessing (EMDR, Shapiro, 1989), a well-known therapeutic approach for the treatment of post traumatic stress disorder (e.g., Lee & Cuijpers, 2013). During EMDR, patients are required to recall the traumatic memory while performing horizontal eye movements. EMDR is argued to change the traumatic (sensory) memory into a more (verbal) declarative memory, while simultaneously reducing patients' emotional arousal and avoidance.

As a result of the suggested association between eye movements and memory in clinical contexts, the past decades have witnessed a growing number of experimental studies on the effects of horizontal eye movements (for a review, see Christman & Propper, 2010). Eye movement experiments typically employ either free recall or recognition memory paradigms and require participants to perform 30 seconds of horizontal eye movements immediately prior to the test phase. According to the alternating hemispheric activation hypothesis (Christman, Garvey, Propper, & Phaneuf, 2003; Propper & Christman, 2008), alternating horizontal eye movements result in the alternating activation of the two brain hemispheres. This activation pattern may lead to increased hemispheric communication, which in turn benefits the retrieval of memories. As strongly right-handed individuals show lower interhemispheric interaction than mixed- and left-handed individuals, the benefits of horizontal saccades are typically more pronounced for strongly right-handers (e.g., Brunyé, Mahoney, Augustyn, & Taylor, 2009; Lyle, Logan, & Roediger, 2008; Lyle, Hanaver-Torrez, Hackländer, & Edlin, 2012).

Consistent with the alternating hemispheric activation hypothesis, the majority of eye movement studies report that horizontal eye movements improve episodic memory retrieval compared to no eye movements, especially for strongly right-handed participants (e.g., Brunyé et al., 2009; Christman et al., 2003; Christman, Propper, & Dion, 2004; Lyle et al., 2008; Lyle & Osborn, 2011; Nieuwenhuis et al., 2013; Parker, Buckley, & Dagnall,

2009; Parker & Dagnall, 2007, 2010, 2012; Parker, Relph, & Dagnall, 2008). Similarly, various studies show that horizontal eye movements improve memory performance compared to vertical eye movements (e.g., Brunyé et al., 2009; Christman et al., 2003; Parker et al., 2009; Parker & Dagnall, 2007, 2012; Parker et al., 2008). The literature is, however, not entirely consistent. First, Lyle et al. (2008) reported that vertical eye movements –similar to horizontal eye movements– improve memory retrieval compared to no eye movements. Second, Samara, Elzinga, Slagter, and Nieuwenhuis (2011) found that the beneficial effect of horizontal eye movements was only present for the recall of emotional stimuli.

Motivated in part by the above mentioned inconsistencies, the skeptics (i.e., the first, third, and sixth author) have recently conducted two pilot studies in which they attempted to replicate the beneficial effect of horizontal eye movements on free recall. The skeptics compared the recall of emotional and neutral study words from Samara et al. (2011) after horizontal and vertical eye movements. In the first study, the skeptics tested 19 strongly right-handed participants in a within-subject design and found no difference in recall performance between the two eye movement conditions. In the second study, the skeptics tested 16 strongly right-handed participants in a between-subject design. In line with the first study, no differences were found between the horizontal and vertical eye movement condition. The skeptics were thus unable to replicate the beneficial effect of horizontal eye movements on free recall performance.

In light of the somewhat inconsistent results in the literature and the additional null results obtained in the two pilot studies, the skeptics invited the proponents (i.e., second and fourth author) to participate in the present adversarial collaboration. Prior to data collection, the adversaries appointed an impartial referee (i.e., the fifth author) and set up an adversarial collaboration agreement. The adversarial collaboration agreement was registered at the OSF before a single participant was tested. The preregistration and the

agreement are available at `http://osf.io/LAyZm/`.

Proposed Experiment and Expectations

The proposed experiment was an attempt to establish whether horizontal eye movements improve episodic memory retrieval. The investigation followed a strictly confirmatory design and relied on preregistered statistical analyses. The adversaries agreed that the proposed design best reflected the prototypical experiment in the field, and that the results were potentially the most compelling to both skeptics and proponents.

Participants were presented with a list of neutral study words for a subsequent free recall test. Prior to recall, participants were requested to perform –depending on the experimental condition– either horizontal, or vertical, or no eye movements (i.e., looking at a central fixation point). The type of eye movement was thus manipulated between subjects. As the effect of eye movement on episodic memory has been reported to be influenced by handedness, we tested only strongly right-handed individuals. The dependent variable of interest was the number of correctly recalled words.

The proponents expected horizontal eye movements to affect recall performance. Specifically, the proponents expected that the number of correctly recalled words (1) was higher in the horizontal than in the no eye movement condition, and (2) was higher in the horizontal than in the vertical eye movement condition. The proponents did not expect the number of correctly recalled words to differ between the vertical and the no eye movement condition. In contrast, the skeptics did not expect horizontal eye movements to affect recall performance. Specifically, the skeptics did not expect the number of correctly recalled words to differ between (1) the horizontal and no eye movement condition, (2) the horizontal and vertical eye movement condition, and (3) the vertical and no eye movement condition.

To demonstrate that the results are not contaminated by unintended peculiarities of

the experimental setting, the skeptics and the proponents also attempted to replicate the well-established associative-priming effect using a lexical decision task (e.g., de Groot, 1984, 1987; Neely, 1976, 1977). The associative-priming task required participants to categorize letter strings as words or nonwords. Each target word was preceded by a prime word that was either an associate of the target (e.g., dog-cat) or was unrelated to the target (e.g., uncle-cat). The dependent variable of interest was the mean response time (RT) for correct responses to target words. Typically, mean correct RTs are shorter for target words preceded by related primes than for target words preceded by unrelated primes. The detailed description of the associative-priming task is available in the adversarial collaboration agreement.

Data Analysis

We believe that in adversarial collaborations it is highly desirable to quantify evidence in favor of the null hypothesis. Moreover, we believe that it is desirable to collect data until the pattern of results is sufficiently clear. As both requirements can be conveniently accomplished within the framework of Bayesian inference, the present experiment relied on hypothesis testing using the Bayes factor (e.g., Berger & Mortera, 1999; Edwards, Lindman, & Savage, 1963; Jeffreys, 1961; Kass & Raftery, 1995; Rouder, Morey, Speckman, & Province, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, 2007; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Wagenmakers et al., 2011, 2012; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009).

The Bayes factor (BF_{01}) is a Bayesian model selection measure that quantifies the probability of the data under the null hypothesis (H_0) relative to the probability of the data under the alternative hypothesis (H_1). The subscript 01 in BF_{01} indicates that we compute the probability of the data under H_0 relative to the probability of the data under H_1. In contrast, the subscript 10 would indicate that we compute the probability of the

data under H_1 relative to the probability of the data under H_0. For instance, $BF_{01} = 10$ indicates that the data are 10 times more likely under H_0 than under H_1. Alternatively, $BF_{01} = \frac{1}{10}$ indicates that the data are 10 times more likely under H_1 than under H_0.

Within the framework of Bayesian inference, the intention with which the data are collected is irrelevant (Berger & Wolpert, 1988; Edwards et al., 1963; Rouder, 2014); hence we can monitor the Bayes factor as the data are collected, and may stop collecting data whenever the evidence is sufficiently compelling. The adversaries agreed to monitor the Bayes factor after each week of data collection and adaptively increase the sample size until a predefined BF has been reached. Specifically, skeptics and proponents set out to test at least 20 participants in each of the three eye movement conditions and agreed to stop testing whenever the Bayes factor for the horizontal eye movement vs. no eye movement condition comparison reflects "strong" evidence for H_0 (i.e., $BF_{01} > 10$) or H_1 (i.e., $BF_{01} < \frac{1}{10}$; see Jeffreys, 1961, for a classification scheme for the Bayes factor). The adversarial collaboration agreement contains the precise specification of the stopping rule.

The present study did not rely on prospective sample size calculations. Nevertheless, Bayesian sample size planning can be very useful in the design stage of experiments. In the Bayesian framework, sample size planning can proceed either based on the expected value of the Bayes factor given a particular sample size or based on the expected sample size that is necessary to reach a predefined Bayes factor. In the first scenario, one generates a series of synthetic data sets with a given effect size and the sample size of interest, compute the Bayes factor for each data set, and obtain the distribution of Bayes factors across the replicated data sets. In the second scenario, one generates a series of synthetic data sets with a given effect size, adding participants to each data set until the desired Bayes factor has been reached, and obtain the distribution of the necessary sample size across the replicated data sets. For planning, it is important to average over all possible experimental results that could be obtained. For inference, however, one only

considers the result that was actually obtained: after the data have been collected, all that matters for the assessment of evidence is the Bayes factor (Wagenmakers, Verhagen, Ly, Bakker, et al., in press). Although Bayes factors provide a continuous measure of evidence, optional stopping using a predefined value of the Bayes factor introduces the possibility of Type I and Type II errors (i.e., misleading evidence; Royall, 2000). In Bayesian hypothesis testing, error rate is not controlled by the researchers; rather it depends on the desired value of the Bayes factor and the minimum sample size formulated in the stopping rule, as well as the population effect size and its prior distribution under H_1 (Sanborn & Hills, 2014; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2014; Royall, 2000).

Sequential hypothesis testing can also be accomplished within the frequentist framework using, for instance, group sequential or adaptive designs (e.g., Lai, Lavori, & Shih, 2012; Proschan, Lan, & Wittes, 2006; see Lakens, 2014 for an introduction for psychologists). Bayesian sequential testing is, however, more flexible because—as opposed to its frequentist counterpart—it does not require researchers to specify the total duration of the data collection period (e.g., Reboussin, DeMets, Kim, & Lan, 2000) or the number of interim analyzes in advance (e.g., Pocock, 1977). Bayesian inference allows investigators to monitor the strength of the evidence continuously over the course of the data collection until a desired Bayes factor has been reached. Note also that, from a frequentist perspective, sequential sampling plans, including the Bayesian ones, all result in biased effect size estimates (e.g., Emerson, Kittelson, & Gillen, 2007; Kruschke, 2013). The focus of the present investigation was, however, not on the estimation of effect size, but on quantifying the evidence for the opinion of the adversaries.

Skeptics and proponents agreed to test the three hypotheses using default unpaired Bayesian t tests as specified by Wetzels et al. (2009). This test relies on the default Jeffreys-Zellner-Siow prior setting, the standard choice for model selection in regression models (Liang, Paulo, Molina, Clyde, & Berger, 2008) and in the t test (Rouder et al.,

2009; Wagenmakers et al., 2011, 2012). The test assumes a Cauchy distribution for the effect size under H_1 with a location parameter of zero and a scale parameter of one (i.e., $\delta \sim \text{Cauchy}(0, 1)$). The Cauchy distribution resembles a standard normal distribution with relatively fat tails, reflecting lack of knowledge about the effect size in a particular paradigm. The Cauchy distribution has been proposed as an objective prior and results in a conservative test.

As the proponents had specific expectations about the direction of the effects (e.g., better recall in the horizontal than in the no eye movement condition), the adversaries used order-restricted (i.e, one-sided) t tests, resulting in a folded Cauchy distribution for effect size that is defined for positive numbers only (i.e., $\delta \sim \text{Cauchy}(0, 1)^+$). Note that neither party expected differences in recall performance between the vertical and the no eye movement condition. The adversaries nevertheless decided to use a one-sided t test because a few studies in the literature reported that—similar to horizontal eye movements—vertical eye movement may also improve episodic memory (e.g., Lyle et al., 2008).

Note that our Bayes factor calculation relies on an underlying model that assumes continuous data. The eye movement data, however, consist of the number of correctly recalled words out of a fixed number of trials (i.e., counts). A more natural description of the data would rely on a Bayesian hierarchical approach (e.g., Farrell & Ludwig, 2008; Gelman & Hill, 2007; Kruschke, 2010; Matzke & Wagenmakers, 2009; Rouder, Lu, Speckman, Sun, & Jiang, 2005) that assumes a binomial distribution to model the recall probability of each participant. In a hierarchical setting, the individual recall probability parameters are then assumed to be drawn from a group-level distribution that describes the between-subject variability of the individual parameters and quantifies the group-level recall probability for each condition. The literature, however, relies exclusively on the familiar t test and ANOVA to assess the effect of eye movement condition on free recall

performance, and therefore we will not discuss the Bayesian hierarchical approach in further detail. We thank one of the reviewers for conducting the appropriate hierarchical analyses and confirming that these yield qualitatively similar, but numerically less extreme effect size estimates than the analyses reported in the present paper.

Methods

The detailed description of the materials and the procedures of the experiment is also available in the adversarial collaboration agreement.

Participants

Participants were recruited from the psychology student pool of the University of Amsterdam. The degree of handedness within this pool of subjects had been assessed with the Edinburgh Handedness Inventory (EHI; Oldfield, 1971) as part of an earlier test battery (i.e., the UvA "testweek"). Handedness scores range from −100 (strongly left) to +100 (strongly right) in steps of 5. Individuals with EHI score equal to or above +80 were considered strongly right-handed and were approached to participate in the experiment.

Skeptics and proponents agreed to exclude the data of two participants: one participant was under the influence of drugs, whereas the other participant failed to provide a valid EHI score. The remaining 79 participants (17 male; mean age 21.22 years; mean EHI 95.06) had normal or corrected-to-normal vision, were native speakers of Dutch, and were not diagnosed with dyslexia. Participation was rewarded with course credits or with €10.

Tasks and Stimuli

The study list for the free recall task consisted of a primacy buffer of three words, 72 experimental words, and a recency buffer of three words. The study words were neutral Dutch words that featured in Zeelenberg, Wagenmakers, and Rotteveel (2006). The

stimulus words are available from the adversarial collaboration agreement. Before the presentation of the first word, a fixation cross appeared in the middle of the screen for 3000 ms. The study words were then presented sequentially in black using lower-case 34 point Arial in the middle of a light-gray display for 2000 ms, with an inter-stimulus interval of 500 ms. The order of word presentation was randomized across participants.

The computerized eye movement task started with a central fixation cross presented against a light-gray display for 3000 ms. In the horizontal and vertical eye movement conditions, participants were instructed to follow a black circle with a diameter of approximately 4° visual angle with their eyes. The circle alternated between the left and right (horizontal eye movements) or between the top and bottom (vertical eye movements) portion of the display for 30 sec. As the circle changed position every 500 ms, participants performed two saccadic eye movements per second. The distance between the left and right position of the circle was the same as the distance between the top and bottom position, namely 27°. In the no eye movement condition, a colored circle was presented at the center of the display. The circle changed color every 500 ms, alternating between blue and red. In all three conditions, the viewing distance from the monitor was approximately 45 cm.

Procedure

Participants were tested individually. Participants were seated behind the computer screen and were given an explanation of the tasks. For the free recall test, participants were explicitly instructed to memorize the presented words for a subsequent memory test. During the eye movement sequence, the experimenter unobtrusively watched participants' eyes in order to ensure that they performed the saccadic (as opposed to smooth persuit) eye movements in the required direction (horizontal, vertical, or fixation), for the required duration (30 sec.), and without accompanying head movements.

Participants were randomly assigned to the three eye movement conditions based on the order of arrival (i.e., Participant 1 was assigned to the horizontal eye movement condition, Participant 2 to the vertical eye movement condition, Participant 3 to the no eye movement condition, Participant 4 to the horizontal eye movement condition, etc.). Participants were then presented with the study list and performed—depending on the eye movement condition—horizontal, vertical, or no eye movements. Next, participants performed a 5-minute paper-and-pencil free recall test.

After a 10-minute break, participants carried out the associative-priming task. Lastly, participants completed an exit interview, inquiring about their age and gender. In addition, participants were asked to indicate whether they were aware of the goal of the experiment, and if so, they were asked to describe what they thought the goal was.

Results

Confirmatory Analyses

Eye movement task. The free recall data are available at `http://osf.io/pXT3M/`. Based on the exclusion criteria specified in the adversarial collaboration agreement, we excluded the free recall data of two participants (one participant in the horizontal and one in the vertical eye movement condition) who correctly described the key hypothesis of the eye movement experiment; the goal was to identify participants who were aware of the hypotheses of the eye movement study and—as a result of their expectations—might have biased the outcome of the free recall task. We also excluded the free recall data of four additional participants (one participant in the horizontal, one in the vertical, and two in the no eye movement condition) who recalled fewer than five items correctly; as described in the adversarial collaboration agreement, skeptics and proponents agreed that participants who were unable to recall at least five out of the 78 words probably did not perform the task seriously. The analyses reported below are based on the data of 25

participants in the horizontal, 24 participants in the vertical, and 24 participants in the no eye movement condition. The sample size of the current experiment closely approximates the median sample size (i.e., 25 participants per condition) used in between-subject free recall studies in the relevant eye movement literature.

The left panel of Figure 1 shows the average number of correctly recalled experimental words in the three eye movement conditions; on average, participants in the horizontal eye movement condition recalled the fewest words and participants in the no eye movement condition recalled the most words. The average number of correctly recalled words was 10.88 (SD = 4.32) in the horizontal, 12.96 (SD = 5.89) in the vertical, and 15.29 (SD = 6.38) in the no eye movement condition. The right panel of Figure 1 shows the posterior distribution of each of the effect sizes. In Bayesian inference, the posterior distribution quantifies the uncertainty about an estimated parameter (i.e., effect size) conditional on the evidence provided by the data. The posterior distributions assign most mass to negative effect sizes. Thus, consistent with the observed data, the posterior distributions for the effect sizes indicate that participants recalled fewer words in the horizontal eye movement condition than either in the vertical or the no eye movement condition and that participants recalled fewer words in the vertical than in the no eye movement condition. Effect size is the largest for the horizontal vs. no eye movement comparison. The horizontal vs. vertical and the vertical vs. no eye movement comparisons resulted in smaller and nearly identical effect size estimates.

As Bayesian inference allows for sequential hypothesis testing, we computed the Bayes factor after each triad of participants. The left panels of Figure 2 show the results of the sequential analyses using one-sided unpaired Bayesian t tests under the assumption of equal variances. The sequential analysis plots show the log Bayes factor as a function of the number of participants per condition; log Bayes factors smaller than zero indicate evidence for H_1, whereas log Bayes factors higher than zero indicate evidence for H_0.

For all three hypotheses, the evidence in favor of H_0 gradually increased as the data accumulated. After testing 73 participants, the Bayes factor indicated that the data are about 15 times more likely under the H_0 of no difference between the horizontal and the no eye movement condition than under H_1 ($BF_{01} = 15.39$).[1] Similarly, the Bayes factor indicated that the data are about 10 times more likely under the H_0 of no difference between the horizontal and the vertical eye movement condition than under H_1 ($BF_{01} = 10.12$). Lastly, the Bayes factor indicated that the data are about 10 times more likely under the H_0 of no difference between the vertical and the no eye movement condition than under H_1 ($BF_{01} = 9.64$). As shown in the right panels of Figure 2, essentially the same results were obtained under the assumption of unequal variances. Unsurprisingly, the frequentist alternatives of the one-sided unpaired tests yielded non-significant results: $t(47) = -2.85$, $p > .99$ for the horizontal vs. no eye movement comparison, $t(47) = -1.41$, $p = .92$ for the horizontal vs. vertical comparison, and $t(46) = -1.32$, $p = .90$ for the vertical vs. no eye movement comparison, assuming equality of variances.

In sum, as anticipated by the skeptics, the Bayes factor indicated strong evidence in favor of H_0 for the horizontal vs. no eye movement as well as the horizontal vs. vertical eye movement comparisons. As expected by both parties, the Bayes factor indicated substantial evidence in favor of H_0 for the vertical vs. no eye movement comparisons.

Associative-priming task. The mean RTs for target words preceded by related primes (493.96 ms, SD = 66.44) were shorter than mean RTs for target words preceded by unrelated primes (527.06 ms, SD = 66.35). The Bayes factor indicated that the data are 528,848,417 times more likely under H_1 than under H_0 (BF_{01} =1.890901E-09). This result supports both parties' expectation and indicates extreme evidence for the presence of the associative-priming effect. The detailed discussion of the results is available in the supplemental materials at `http://osf.io/pXT3M/`.

Exploratory Analyses

This section presents the results of a series of exploratory analyses of the free recall data. First, we examine the sensitivity of the conclusions with respect to the prior setting used in the confirmatory analyses. Second, we probe the robustness of the conclusions to the preregistered outlier treatment. Third, we investigate the effect of eye movements on the number of incorrectly recalled words. Note that all analyses presented in this section are post-hoc, and therefore are not preregistered or described in the adversarial collaboration agreement.

Prior Distribution of Effect Size. Here we present the results of a series of analyses aimed at exploring the robustness of the conclusions with respect to the prior setting used for the analysis of the free recall data. In order to minimize the role of subjective expectations, the confirmatory analyses assumed the default Cauchy$(0, 1)^+$ prior for effect size. The choice of the Cauchy prior may nevertheless be disputed; we might just as well have used a prior that is informed by the eye-movement literature or a prior that assumes smaller variability in effect size than the default Cauchy distribution. Especially the latter possibility warrants further investigation as Bayes factors are sensitive to the variability of the prior distribution (e.g., Bartlett, 1957; Liu & Aitkin, 2008; Vanpaemel, 2010). Specifically, wide prior distributions define highly complex models (i.e., models that can generate a wide range of predictions), resulting in Bayes factors that support H_0. Thus, highly uninformative prior distributions yield Bayes factors that lend infinite support for H_0 (Jeffreys, 1961).

Here we investigate the extent to which the variability of the prior distribution of effect size influences the Bayes factor. We replaced the Cauchy prior on effect size with a zero centered normal prior and varied the standard deviation (SD) from 0 to 2, creating progressively more spread out—uninformative—priors. As we are concerned with

one-sided tests, we used a normal prior that is defined for positive numbers only (i.e., $\delta \sim \text{Normal}(0, \text{SD})^{+}$). The analyses reported in this section are based on the data of 25 participants in the horizontal, 24 participants in the vertical, and 24 participants in the no eye movement condition using the outlier treatment specified in the adversarial collaboration agreement.

Figure 3 shows changes in the log Bayes factor as a function of the standard deviation of the normal prior on effect size. The black triangle corresponds to the Bayes factor computed with the standard normal prior—the so-called unit information prior—on effect size (i.e., $\delta \sim \text{Normal}(0, 1)$). As before, log Bayes factors smaller than zero indicate evidence for H_1, whereas log Bayes factors higher than zero indicate evidence for H_0. Two aspects of the results are noteworthy. First, as the standard deviation of the normal prior increases (i.e., prior becomes progressively wider), the Bayes factor increasingly favors H_0. As mentioned above, this result reflects a typical aspect of Bayesian hypothesis testing. Second, the log Bayes factor is never smaller than zero. This result indicates that the Bayes factor never favors H_1 over H_0 regardless of the variability of the prior distribution. Even under the prior setting that maximally supports H_1 (i.e., standard deviation very close to zero), the log Bayes factor is only around 0, indicating perfectly ambiguous evidence. This finding is not surprising; mean recall was highest in the no eye movement condition and lowest in the horizontal eye movement condition, a result that contradicts the order-restriction specified for the one-sided t test.

The results of the robustness analyses indicated that the Bayes factor, as expected, varied as a function of the standard deviation of the prior distribution of the effect size. Although the strength of the support for H_0 varied as a function of the prior setting, the Bayes factor always favored H_0 over H_1 regardless of the variability of the prior.

Outlier Treatment. In this section we present a series of analyses aimed at probing the robustness of the conclusions to the outlier treatment and the corresponding exclusion

criterion specified in the adversarial collaboration agreement. Specifically, prompted by a reviewer we decided to explore the exclusion criterion "recall score more extreme than average recall $+/-3 \times SD$". This new exclusion criterion did not result in the removal of any participants; consequently, the analyses reported in this section are based on the data of 26 participants in the horizontal, 25 participants in the vertical, and 26 participants in the no eye movement condition.

On average, participants in the horizontal eye movement condition recalled the fewest words and participants in the no eye movement condition recalled the most words. The average number of correctly recalled words was 10.5 (SD = 4.66) in the horizontal, 12.44 (SD = 6.31) in the vertical, and 14.27 (SD = 7.10) in the no eye movement condition. We used default Bayesian unpaired one-sided t tests under the assumption of equal variances to quantify the evidence that the data provide for the preregistered hypotheses. The Bayes factor indicated that the data are about 14 times more likely under the H_0 of no difference between the horizontal and the vertical eye movement condition than under H_1 ($BF_{01} = 13.81$). Similarly, the Bayes factor indicated that the data are about 10 times more likely under the H_0 of no difference between the horizontal and the vertical eye movement condition than under H_1 ($BF_{01} = 9.68$). Lastly, the Bayes factor indicated that the data are about nine times more likely under the H_0 of no difference between the vertical and the no eye movement condition than under H_1 ($BF_{01} = 8.54$). Essentially the same results were obtained under the assumption of unequal variances: $BF_{01} = 13.64$ for the horizontal vs. no eye movement comparison, $BF_{01} = 9.63$ for the horizontal vs. vertical eye movement comparison, and $BF_{01} = 8.61$ for the vertical vs. no eye movement comparison. In sum, the results of the robustness analyses indicated that the conclusions from the free recall data are unaffected by the choice of outlier treatment.

False Alarms. Eye movement studies that rely on recognition memory paradigms often report a reduction in the number of incorrectly recalled words (i.e., false alarms)

after horizontal eye movements relative to vertical or no eye movements (e.g., Lyle et al., 2012; Parker & Dagnall, 2007; Parker et al., 2009). In free recall paradigms, the beneficial effect of horizontal eye movements on false alarms is less conclusive. A few studies reported a reduction in false alarms after horizontal eye movements relative to no eye movements (e.g., Christman et al., 2004; Lyle et al., 2008). Other studies, however, failed to find differences in the number of false alarms as a function of eye movement condition (Nieuwenhuis et al., 2013) or omitted the analysis of false alarms altogether (e.g., Christman, Propper, & Brown, 2006; Parker & Dagnall, 2010; Parker et al., 2008; Samara et al., 2011).

For the sake of comprehensiveness, here we present a series of analyses aimed at exploring the effect of eye movements on the number of false alarms. We used default Bayesian unpaired one-sided t tests under the assumption of equal variances to asses whether (1) horizontal eye movements reduce the number of false alarms relative to no eye movements; (2) horizontal eye movements reduce the number of false alarms relative to vertical eye movements; and (3) vertical eye movements reduce the number of false alarms relative to no eye movements. The analyses reported in this section are based on the data of 25 participants in the horizontal, 24 participants in the vertical, and 24 participants in the no eye movement condition using the outlier treatment specified in the adversarial collaboration agreement.

The average number of false alarms was 2.48 (SD = 2.65) in the horizontal, 4.08 (SD = 5.27) in the vertical, and 3.46 (SD = 2.40) in the no eye movement condition. After 73 participants, the Bayes factor indicated that the data are about as likely under the H_0 of no difference between the horizontal and the no eye movement condition as under H_1 ($BF_{01} = 1.16$). Similarly, the Bayes factor indicated that the data are about as likely under the H_0 of no difference between the horizontal and the vertical eye movement condition as under H_1 ($BF_{01} = 1.15$). Lastly, the Bayes factor indicated that the data are

about seven times more likely under the H_0 of no difference between the vertical and the no eye movement condition than under H_1 ($BF_{01} = 6.51$). Essentially the same results were obtained under the assumption of unequal variances: $BF_{01} = 1.19$ for the horizontal vs. no eye movement comparison, $BF_{01} = 1.27$ for the horizontal vs. vertical eye movement comparison, and $BF_{01} = 6.66$ for the vertical vs. no eye movement comparison.

The data in the vertical eye movement condition featured an outlier of 27 false alarms. After removing this data point, the mean number of false alarms decreased to 3.09 (SD = 2.02) in the vertical eye movement condition. The corresponding Bayes factors were $BF_{01} = 2.03$ for the horizontal vs. vertical eye movement comparison and $BF_{01} = 2.78$ for the vertical vs. no eye movement comparison (equal variances assumed).

In sum, the analysis of the number of false alarms indicated that the evidence for the beneficial effect of horizontal eye movements relative to vertical and no eye movements is almost perfectly ambiguous. The comparison of the vertical and no eye movement conditions yielded moderate evidence for H_0. The evidence for H_0, however, decreased substantially after removing a single outlying observation. These results indicate that the false alarm data are not sufficiently diagnostic to discriminate between H_0 and H_1.

Discussion

Adversarial collaboration has been repeatedly advocated as a constructive method of scientific conflict resolution (Hofstee, 1984; Kahneman, 2003; Latham et al., 1988; Mellers et al., 2001). We believe that adversarial collaborations—especially when coupled with preregistration—may also remedy a number of factors that contributed to the crisis of confidence in psychological science and increase the transparency of scientific communication (see also Koole & Lakens, 2012; Wagenmakers et al., 2011). The present paper therefore introduced the notion of preregistered adversarial collaboration, a novel variant of scientific conflict resolution. The proposed approach combines the features of

adversarial collaboration and purely confirmatory research (Wagenmakers et al., 2012).

We illustrated the use of preregistered adversarial collaboration with a joint proponent-skeptic investigation on the effect of horizontal eye movements on episodic memory performance. The rules of the collaboration were as follows. First, the adversaries reached consensus on an optimal research design. Specifically, the adversaries agreed to manipulate the type of eye movement between subjects: participants were requested to perform either horizontal, or vertical, or no eye movements prior to the recall of the study list. Second, the two parties formulated their expectations and agreed to submit the findings to an academic journal whether or not those expectations are supported by the data. Third, the adversaries appointed an impartial referee whose task was to oversee the collaboration. Lastly, but importantly, the two parties set up a publicly available adversarial collaboration agreement that described the proposed design and all foreseeable aspects of the data analysis. The adversarial collaboration agreement was registered at the OSF before a single participant was tested. The adversarial collaboration agreement presented here may serve as a blueprint for future work.

As expected by the skeptics, the Bayes factor indicated strong evidence in favor of H_0 for the horizontal eye movement vs. no eye movement as well as for the horizontal eye movement vs. vertical eye movement comparisons. As expected by both parties, the Bayes factor indicated substantial evidence in favor of H_0 for the vertical eye movement vs. no eye movement comparison. Lastly, the results of the associative-priming task supported both parties' expectation and indicated extreme evidence for the presence of an associative-priming effect. In what follows, the skeptics and the proponents will present their own perspectives on the results of the experiment and the process of the joint research effort.

Discussion by Skeptics

Reflection on the results. The results clearly supported our expectations: horizontal eye movements did not improve free recall performance in the present experiment. Despite our best efforts to carry out a prototypical experiment, the present study—and our two pilot studies—thus failed to replicate the seemingly well-established effect of bilateral eye movements on episodic memory and failed to find evidence for the predictions of the alternating hemispheric activation hypothesis (Christman et al., 2003; Propper & Christman, 2008).

Our failure to replicate may, of course, simply be due to chance; even if the effect under scrutiny truly exists, a certain number of replication attempts are necessarily doomed to be unsuccessful (e.g., Francis, 2013). Note, however, that our two pilot studies also yielded null results. We propose therefore that the conflicting findings may reflect mechanisms that are related to (1) statistical problems in the literature; (2) prevailing research practices in psychology; and (3) methodological shortcomings of the prototypical research design.

On the statistical side, we believe that the effect of horizontal eye movements on episodic memory may be overestimated as a result of the statistical problems associated with p value-based null hypothesis testing. A well-known problem of frequentist hypothesis testing is that p values overstate evidence against H_0 (Berger & Delampady, 1987; Edwards et al., 1963; Johnson, 2013; Sellke, Bayarri, & Berger, 2001). Wetzels et al. (2011) showed that 70% of the p values from t tests in experimental psychology that fall between .01 and .05 correspond to default Bayes factors that indicate that the data are no more than three times more likely under H_1 than under H_0. This suggests that a number of "significant" findings in the eye movement literature (e.g., Brunyé et al., 2009; Lyle et al., 2008; Samara et al., 2011) may in fact reflect negligible effects that are "not worth more than a bare mention" (Jeffreys, 1961). We believe that adopting a more strict

significance level, say $\alpha = 0.01$, would not remedy this problem; because the p value decreases with increasing sample size, researchers could simply increase the sample size of their experiments to adapt to the more strict significance level (Wetzels et al., 2011). The source of the discrepancy between p values and Bayes factors is that the p value only considers the plausibility of the data under H_0. The p value therefore ignores the possibility that the data may be just as extreme—or even more extreme—under H_1 (Berkson, 1938; Wagenmakers, Verhagen, Ly, Matzke, et al., in press). The present paper therefore advocates the use of Bayesian hypothesis testing with default Bayes factors.

Although it is likely that the eye movement literature is biased by the statistical peculiarities of p values, the results of the present experiment cannot be explained purely in terms of differences in statistical framework. The Bayesian conclusions were corroborated with the results of p value-based hypothesis tests. In fact, participants in the horizontal eye movement condition recalled on average the fewest words, a result that contradicts most—if not all—reported findings in the eye movement literature.

We therefore argue that the conflicting results may partly reflect bias and the use of questionable research practices, both of which can distort the literature. That is, the beneficial effect of horizontal eye movements on free recall may seem more established than it actually is, due to publication bias and the file-drawer problem (Rosenthal, 1979; Greenwald, 1975). Hindsight bias and positive confirmation bias during the interpretation of the data may likewise contribute to the unbalanced literature by fueling the use of questionable research practices (QRP). QRPs may include optional stopping (i.e., collecting data until the p value reaches a desired significance criterion), selectively reporting results from experimental conditions and dependent variables that produce significant effects, hypothesizing after the results are known (HARKing; Kerr, 1998), and the use of post-hoc exclusion criteria, such as arbitrary handedness cut-off scores. For instance, the following investigations all used different criteria for classifying participants

as strongly right-handed: Brunyé et al. (2009) used EHI > median, Christman et al. (2004, Experiment 1) used EHI ≥ median, Christman et al. (2004, Experiment 2) used EHI ≥ 75, and Lyle and Osborn (2011) used EHI ≥ 80. The present paper therefore emphasizes the importance of preregistration and the strict separation of confirmatory and exploratory research (see also de Groot, 1956/2014).

Lastly, on the methodological side, we argue that limitations of the prototypical research design may contribute to the conflicting findings. In the present study, as in most eye movement studies, the experimenter was not blind to participants' eye movement condition. The expectations of the experimenter may have unintentionally influenced the outcome of the study by, say, selectively increasing participants' motivation in a given eye movement condition (Rosenthal, 1976). In the present study, the data were collected by the skeptics. Despite our best efforts, our expectations might have been subtly communicated to the participants and have contributed to the null finding in the present experiment and in our two pilot studies. The possibility of the experimenter bias as an explanation for our results warrants further investigation. Note however that if the experimenter's expectation can indeed eliminate or even reverse the effect of bilateral eye movement on free recall, the phenomenon is more fragile than suggested by the literature, a possibility that may explain the present failure to replicate.

Reflection on the process. Preregistered adversarial collaboration is a labor-intensive undertaking that requires more planning and anticipation than carrying out standard research. Prior to data collection, the adversaries are required to reach consensus on an experimental design and have to anticipate and document—as far as possible—all foreseeable aspects of the data collection and the data analysis. We believe, however, that the advantages of the proposed approach outweigh the disadvantages, as the initial effort involved in setting up the joint research pays off in numerous ways. By critically evaluating and attempting to anticipate all aspects of the research effort, the two parties

capitalize on expert knowledge and maximize the probability that the proposed experiment resolves the disagreement. Moreover, the public disclosure of the experimental procedures and statistical analyses secures the purely confirmatory nature of the research and increases the transparency of the investigation.

Note that preregistration of the proposed experiment does not mean that all aspects of the research effort are carved in stone. If both parties agree, the adversarial collaboration agreement may be amended to account for unexpected events during data collection. For instance, as documented in the present adversarial collaboration, we agreed to modify the stopping rule and our strategy for participant recruitment during data collection (see amendment to the adversarial collaboration agreement on the OSF and footnote 1). Similarly, preregistration of the data analysis does not mean that investigators cannot follow up interesting patterns in the data or—as demonstrated here—investigate the robustness of the conclusions. We believe that exploratory research plays an essential role in science; it generates new testable hypotheses and facilitates scientific progress. We also believe, however, that researchers should explicitly acknowledge which results are based on explorations and which results are based on strictly confirmatory analyses.

In sum, setting up preregistered joint research requires more effort on behalf of the investigators than carrying out standard research. We believe, however, that the additional work is a small price to pay for the possibility of constructive conflict resolution and a great increase in transparency. We hope that preregistered adversarial collaboration—or some other variant of confirmatory joint research—will in the near future become the rule rather than the exception for settling scientific disputes in psychology. In light of the rather heated debates in our discipline, there is certainly room for improvement.

Discussion by Proponents

Reflection on the results. We were surprised by these results. In a previous study, we found a beneficial effect of horizontal eye movements on recall of emotional words but not neutral words (Samara et al., 2011). However, the null effect for neutral words may have been due to the small sample size ($N = 14$) and/or the relative long period between the horizontal eye movements and subsequent recall test due to an intermittent baseline EEG recording; in a subsequent study, using a much larger sample and no intermittent EEG recording, we did replicate the effect (Nieuwenhuis et al., 2013, Experiment 1). In additional experiments we found a similar beneficial effect on word recall of alternating (vs. simultaneous) left-right tactile but not auditory stimulation, a pattern of results predicted by the alternating hemispheric activation hypothesis (Christman et al., 2003; Propper & Christman, 2008). These and other studies (Propper & Christman, 2008) used procedures and stimulus material that were similar to those used in the current study. In addition, the current study only included consistently right-handed individuals as the effect of horizontal eye movements on memory is present in strong left- and right-handers but not in mixed-handers (Lyle et al., 2008, 2012). It is thus surprising that in the current study, previously reported positive effects of horizontal eye movements on memory performance were not replicated.

So how can we account for the current non-replication? As the skeptics suggest, the non-replication might be a false negative. Or it may be due to experimenter bias (Rosenthal & Rubin, 1978). To rule out this latter possibility, experimenters in future studies will have to be blind to the condition to which a participant is assigned. Here, we consider in more detail another explanation offered by the skeptics: the possibility that researchers selectively report positive studies or analyses, or use any of several questionable strategies (e.g., optional stopping; try different contrasts) for producing a significant effect of horizontal eye movements. To investigate this possibility we conducted

a *p*-curve analysis (Simonsohn, Nelson, & Simmons, 2014). That is, we plotted the distribution of statistically significant p values ($< .05$) reported in studies on the beneficial effects of horizontal eye movements on memory and examined the form of the distribution. As Simonsohn and colleagues argue, "only right-skewed *p*-curves, those with more low (e.g., .01s) than high (e.g., .04s) significant p values, are diagnostic of evidential value. *P*-curves that are not right-skewed suggest that the set of findings lacks evidential value, and *p*-curves that are left-skewed suggest the presence of intense *p*-hacking" (i.e. obtaining statistically significant results using QRPs).

For this analysis, we selected all studies that examined the effects of 30 seconds of horizontal eye movements (relative to a control condition) on explicit memory in consistently-handed healthy individuals. The steps involved in the selection of p values that meet these selection criteria are documented in the recommended *p*-curve disclosure table (cf. Simonsohn et al., 2014) available in the supplemental materials. Figure 4 shows the results of the *p*-curve analysis based on these p values. As can be seen in this figure, the *p*-curve is significantly right-skewed, $\chi^2(36) = 102.33$, $p < .0001$, indicating that these studies do contain evidential value. This means that we can rule out *p*-hacking as the sole explanation for the reported effects of horizontal eye movements. As Simonsohn and colleagues show, with a sample size of ~ 20 p values, it is virtually impossible for *p*-curve analysis to indicate that the sample contains evidential value when in fact the studies were intensely *p*-hacked. Nevertheless, it is worth noting that there is an uptick in the *p*-curve at .05 (test for left skew: $\chi^2(36) = 28.23$, $p = .82$). A *p*-curve is markedly right-skewed when an effect is real but only mildly left-skewed when a finding is *p*-hacked. So Simonsohn and colleagues acknowledge that if a set of findings combines true effects with nonexistent ones, the *p*-curve will usually not detect the latter. Thus, the *p*-curve analysis suggests that the effect of horizontal eye movements on explicit memory is a true effect, but leaves open the possibility that some of the significant findings were *p*-hacked.

The analysis yielded two other noteworthy findings. First, of the 18 p values that were selected for the p-curve analysis, 11 were $< .025$, and 7 of these 11 more significant p values were published by one group (i.e., Parker, Dagnall, and colleagues). Indeed, altogether only 5 different research groups have contributed to the literature examined here. It is thus important that more laboratories will replicate the effect. Second, in the current study, effects of horizontal eye movements on recall were examined. Therefore, we asked whether there was a difference in p values between studies using recall and studies using recognition tests, as it is possible that horizontal eye movements affect one type of memory more strongly than the other. This was not the case: of the 11 p values $< .025$, 5 reflected recall tests and 6 reflected recognition tests. Of the 7 significant p values $> .025$, 4 were based on recall tests, 3 on recognition tests.

Considering the empirical results and the p-curve analysis reported here, did the present adversarial collaboration resolve the disagreement between the skeptics and the proponents? No; the skeptics are probably no less skeptical, and we, the proponents, are not convinced by a single failure to replicate, especially given the results of the p-curve analysis. However, we have become more cautious about the conclusions that can be drawn from the studies reported so far, and will follow the further development of this field of research with a critical eye. It is important to note that although several authors have speculated about a link between this memory literature and a more clinical literature suggesting that eye movements reduce the vividness and distress associated with emotional autobiographical memories, we do not believe that the current results should lead researchers to call into question those clinical findings. A recent meta-analysis has found significant evidence that eye movements affect the processing of distressing memories in eye-movement desensitization and reprocessing (EMDR) therapy (moderate effect size) and in non-therapy contexts (large effect size; Lee & Cuijpers, 2013).

Reflection on the process. Although our adversarial collaboration has not resolved the debate, it has generated new testable ideas and has brought the two parties slightly closer by demonstrating that the beneficial effect of bilateral eye movements on episodic memory is not unequivocal. We recommend that other researchers in this field use similar strict methods in future studies, and emphasize the importance of reporting non-replications.

Discussion by Referee

An impartial referee has been involved in the adversarial collaboration throughout the course of the process. The referee was asked to settle any dispute between parties that might arise with regard to issues not specified in the contract. That did not happen. The parties agreed on the "Adversarial Collaboration Agreement" contract without the need for a referee. The referee received weekly updates during data collection and observed that the parties were able to solve issues not specified in the contract, e.g., the required number of participants or outlier/exclusion criteria, on their own. Finally, and most importantly, the parties agreed upon the outcome of the adversarial collaboration. The results that emerged from this adversarial collaboration show that horizontal eye movements did not improve free recall. Game over and done with? It seems not to be the case. The results are clearly in support of the skeptics' expectations. However, while accepting the negative findings and acknowledging the benefits of preregistered adversarial collaboration, the proponents are not convinced by a single failure to replicate, especially given the results of the p-curve analysis.

Thus, we have to conclude that the adversarial collaboration could not settle the empirical debate conclusively: despite the highly diagnostic outcome of the experiment, the proponents are still convinced that the effect is real. In hindsight, this result was to be expected. A single experiment, even when pre-registered and conducted in the framework

of an adversarial collaboration, may not provide sufficient evidence to overturn an opinion that was shaped over the course of many years. In this regard, adversarial collaborations can only be a first step in a larger research program. Even a statistically compelling result (as obtained in the present collaboration) may be insufficient to overcome a long-held belief. A definite answer on the relationship between eye movements and memory must await a series of replications by skeptics as well as proponents, performed in independent labs, using purely confirmatory research designs and statistical analyses.

Alternative Approaches and Future Directions

The goal of this paper was to establish a collaboration between proponents and skeptics and investigate the effect of eye movements on memory in a purely confirmatory setting that is uncontaminated by QRPs. The adversaries attempted to extract the common features of published eye movement studies in order to design a prototypical experiment in which all conditions are optimal for observing the effect. The present approach, however, is only one of many avenues that can be taken to probe the existence of the eye movement effect on memory. An evident alternative is the preregistered direct replication of a landmark experiment. We see merits in both approaches.

Direct replications are crucial for establishing the existence of the phenomenon under scrutiny: direct replications can identify false positives, may facilitate the identification of factors that moderate the (size of the) effect, and can generate skepticism about the existence of seemingly well-established scientific findings (Nosek & Lakens, 2014; Pashler & Harris, 2012). In direct replications, however, any result could be attributed to idiosyncrasies of the experimental design at hand. Replication attempts of eye movement studies are particularly prone to this problem because different labs often rely on subtly different versions of the same basic paradigm (e.g., different stimulus type and study list length).

We believe that a single study, whether a direct replication or a novel prototypical experiment, can only be a first step in a more systematic and large-scale research effort that examines the existence and the boundary conditions of the eye movement effect on memory. As suggested by one of the reviewers, future research should focus on the possible moderating effect of stimulus type (e.g., neutral vs. emotional), study list length, the duration of the retention interval, and the frequency, duration, and spatial extent of the eye movement sequence. Moreover, as pointed out both by skeptics and proponents, future work should examine the possibility of experimenter bias as an explanation for the current non-replication, and may consider quantitative monitoring of the eye movement sequence. Such large-scale replication effort would require collaboration between skeptics and proponents from various labs, and would ideally start with direct replications of a number of landmark experiments, followed by a tree of preregistered studies based on "what-next" contingencies that systematically vary factors that might influence the eye movement effect on memory.

There are several hypotheses about the neural mechanisms underlying eye movement-induced memory improvements, such as the alternating hemispheric activation hypothesis (Christman et al., 2003; Propper & Christman, 2008). Yet, so far, brain evidence is scarce. Although the present replication attempt failed to provide evidence for the predictions of the alternating hemispheric activation hypothesis, future work should also include measures of brain activity, as it remains unclear precisely how eye movements may affect memory in the first place. For the time being, we believe that our adversarial collaboration has generated new testable ideas that may shed further light on the relationship between eye movements and episodic memory and we hope that the present work will trigger a more cautious attitude towards the conclusions that can be drawn from the literature.

References

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554.

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, *44*, 533–534.

Bateman, I., Kahneman, D., Munro, A., Starmer, C., & Sugden, R. (2005). Testing competing models of loss aversion: An adversarial collaboration. *Journal of Public Economics*, *89*, 1561–1580.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Berger, J. O., & Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, *94*, 542–554.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle*. Institute of Mathematical Statistics.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33*, 526–536.

Brunyé, T. T., Mahoney, C. R., Augustyn, J. S., & Taylor, H. A. (2009). Horizontal saccadic eye movements enhance the retrieval of landmark shape and location information. *Brain and Cognition*, *70*, 279–288.

Cadsby, C. B., Croson, R., Marks, M., & Maynes, E. (2008). Step return versus net reward in the voluntary provision of a threshold public good: An adversarial collaboration. *Public Choice*, *135*, 277–289.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, *49*, 609–610.

Chambers, C. D., Munafo, M., & et al. (2013). *Trust in science would be improved by study pre-registration.* Retrieved from `http://www.theguardian.com/science/`

blog/2013/jun/05/trust-in-science-study-pre-registration

Christman, S. D., Garvey, K. J., Propper, R. E., & Phaneuf, K. A. (2003). Bilateral eye movements enhance the retrieval of episodic memories. *Neuropsychology*, *17*, 221–229.

Christman, S. D., & Propper, R. E. (2010). An interhemispheric basis for episodic memory: Effects of handedness and bilateral eye movements. In G. Davies & D. Wright (Eds.), *Current issues in applied memory* (p. 185-205). London, UK: Psychology Press.

Christman, S. D., Propper, R. E., & Brown, T. J. (2006). Increased interhemispheric interaction is associated with earlier offset of childhood amnesia. *Neuropsychology*, *20*, 336-345.

Christman, S. D., Propper, R. E., & Dion, A. (2004). Increased interhemispheric interaction is associated with decreased false memories in a verbal converging semantic associates paradigm. *Brain and Cognition*, *56*, 313–319.

de Groot, A. D. (1956/2014). The meaning of "significance" for different types of research. Translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas. *Acta Psychologica*, *148*, 188-194.

de Groot, A. D. (1969). *Methodology: Foundations of inference and research in the behavioral sciences.* The Hague, The Netherlands: Mouton.

de Groot, A. M. B. (1984). Primed lexical decision: Combined effects of the proportion of related prime–target pairs and the stimulus–onset asynchrony of prime and target. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *36*, 253–280.

de Groot, A. M. B. (1987). The priming of word associations: A levels-of-processing approach. *The Quarterly Journal of Experimental Psychology Section A: Human*

Experimental Psychology, *39*, 721–756.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*, e29081.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Emerson, S. S., Kittelson, J. M., & Gillen, D. L. (2007). Frequentist evaluation of group sequential clinical trial designs. *Statistics in Medicine*, *26*, 5047–5080.

Farrell, S., & Ludwig, C. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic Bulletin & Review*, *15*, 1209–1217.

Francis, G. (2013). Publication bias in "Red, rank, and romance in women viewing men" by Elliot et al. (2010). *Journal of Experimental Psychology: General*, *142*, 292–296.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gilovich, T., Medvec, V. H., & Kahneman, D. (1998). Varieties of regret: A debate and partial resolution. *Psychological Review*, *105*, 602–605.

Goldacre, B. (2009). *Bad science*. London, UK: Fourth Estate.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Hofstee, W. K. B. (1984). Methodological decision rules as research policies: A betting reconstruction of empirical research. *Acta Psychologica*, *56*, 93–109.

Hunter, J. E. (2001). The desperate need for replications. *Journal of Consumer Research*, *28*, 149–158.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of

questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, *110*, 19313–19317.

Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist*, *58*, 723.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*, 608–614.

Kruschke, J. K. (2010). *Doing Bayesian data analysis: A tutorial introduction with R and BUGS*. Burlington, MA: Academic Press.

Kruschke, J. K. (2013). *Doing Bayesian data analysis: P values, Bayes factors, credible intervals, precision.* Retrieved from `http://doingbayesiandataanalysis.blogspot.de/2013/11`

Lai, T., Lavori, P., & Shih, M.-C. (2012). Adaptive trial designs. *Annual Review of Pharmacology and Toxicology*, *52*, 101–110.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*.

Latham, G. P., Erez, M., & Locke, E. A. (1988). Resolving scientific disputes by the joint design of crucial experiments by the antagonists: Application to the Erez–Latham dispute regarding participation in goal setting. *Journal of Applied Psychology*, *73*, 753–772.

Lee, C. W., & Cuijpers, P. (2013). A meta-analysis of the contribution of eye movements

in processing emotional memories. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*, 231.239.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375.

Lyle, K. B., Hanaver-Torrez, S. D., Hackländer, R. P., & Edlin, J. M. (2012). Consistency of handedness, regardless of direction, predicts baseline memory accuracy and potential for memory enhancement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 187-193.

Lyle, K. B., Logan, J. M., & Roediger, H. L. (2008). Eye movements enhance memory for individuals who are strongly right–handed and harm it for individuals who are not. *Psychonomic Bulletin & Review*, *15*, 515–520.

Lyle, K. B., & Osborn, A. E. (2011). Inconsistent handedness and saccade execution benefit face memory without affecting interhemishperic interction. *Memory*, *19*, 613–624.

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex–Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*, 798–817.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*, 269–275.

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, *4*, 648–654.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of

inhibitionless spreading activation and limited–capacity attention. *Journal of Experimental Psychology: General*, *106*, 226–254.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.

Nier, J. A., & Campbell, S. D. (2012). Two outsiders view on feminism and evolutionary psychology: An opportune time for adversarial collaboration. *Sex Roles*, 1–4.

Nieuwenhuis, S., Elzinga, B. M., Ras, P. H., Berends, F., Duijs, P., Samara, Z., & Slagter, H. A. (2013). Bilateral saccadic eye movements and tactile stimulation, but not auditory stimulation, enhance memory retrieval. *Brain and Cognition*, *81*, 52–56.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh Inventory. *Neuropsychologia*, *9*, 97–113.

Parker, A., Buckley, S., & Dagnall, N. (2009). Reduced misinformation effects following saccadic bilateral eye movements. *Brain and Cognition*, *69*, 89–97.

Parker, A., & Dagnall, N. (2007). Effects of bilateral eye movements on gist based false recognition in the DRM paradigm. *Brain and Cognition*, *63*, 221–225.

Parker, A., & Dagnall, N. (2010). Effects of handedness and saccadic bilateral eye movements on components of autobiographical recollection. *Brain and Cognition*, *73*, 93–101.

Parker, A., & Dagnall, N. (2012). Effects of saccadic bilateral eye movements on memory in children and adults: An exploratory study. *Brain and Cognition*, *78*, 238–247.

Parker, A., Relph, S., & Dagnall, N. (2008). Effects of bilateral eye movements on the

retrieval of item, associative, and contextual information. *Neuropsychology*, *22*, 136.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*, 191–199.

Propper, R. E., & Christman, S. D. (2008). Interhemispheric interaction and saccadic horizontal eye movements. Implications for episodic memory, EMDR, and PTSD. *Journal of EMDR Practice and Research*, *2*, 269–281.

Proschan, M., Lan, K., & Wittes, J. (2006). *Statistical monitoring of clinical trials: A unified approach*. Springer.

Reboussin, D. M., DeMets, D. L., Kim, K. M., & Lan, G. K. (2000). Computations for group sequential boundaries using the Lan-DeMets spending function method. *Controlled Clinical Trials*, *21*, 190–207.

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science*, *7*, 411–426.

Rosenthal, R. (1976). *Experimenter effects in behavioral research*. New York, NY: Irvington.

Rosenthal, R. (1979). An introduction to the file drawer problem. *Psychological Bulletin*, *86*, 683–641.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, *1*, 377–386.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*, 195–223.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Royall, R. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, *95*, 760–768.

Samara, Z., Elzinga, B. M., Slagter, H. A., & Nieuwenhuis, S. (2011). Do horizontal saccadic eye movements increase interhemispheric coherence? Investigation of a hypothesized neural mechanism underlying EMDR. *Frontiers in Psychiatry*, *2*, 1–7.

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, *21*, 283–300.

Schlitz, M., Wiseman, R., Watt, C., & Radin, D. (2006). Of two minds: Sceptic–proponent collaboration within parapsychology. *British Journal of Psychology*, *97*, 313–322.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2014). Sequentail hypothesis testing with Bayes factors: A practical and powerful approach. *Manuscript submitted for publication*.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, *55*, 62–71.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., ... Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS ONE*, *8*, e56515.

Shapiro, F. (1989). Eye movement desensitization: A new treatment for post-traumatic

stress disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, *20*, 211–217.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *PsychonomicBulletin & Review*, *14*, 779–804.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M., Matzke, D., ... Morey, R. (in press). A power fallacy. *Behavior Research Methods*.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. (in press). The need for Bayesian hypothesis testing in psychologiacal research. In S. Lilienfeld & I. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions.* John Wiley and Sons.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, 426–432.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291–298.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin &Review*, *16*, 752–760.

Wiseman, R., & Schlitz, M. (1997). Experimenter effects and the remote detecting of staring. *Journal of Parapsychology*, *61*, 197–207.

Wiseman, R., & Schlitz, M. (1998). Replication of experimenter effects and the remote detecting of staring. *Proceedings of the 12nd Annual Convention of the Parapsychological Association*, 471–479.

Wolfe, J. M. (2013). Registered reports and replications in Attention, Perception, & Psychophysics. *Attention, Perception, & Psychophysics*, *75*, 781–783.

Zeelenberg, R., Wagenmakers, E.-J., & Rotteveel, M. (2006). The impact of emotion on perception: Bias or enhanced processing? *Psychological Science*, *17*, 287–291.

Footnotes

[1]After five weeks of data collection, BF_{01} was above 10 for the horizontal eye movements vs. no eye movement comparison. The adversaries, however, agreed to continue testing for an additional week in order to obtain compelling evidence also for the horizontal vs. vertical eye movements and the vertical vs. no eye movement comparisons. For the amendment to the adversarial collaboration agreement that documents this decision, see the OSF at `http://osf.io/pXT3M/`.

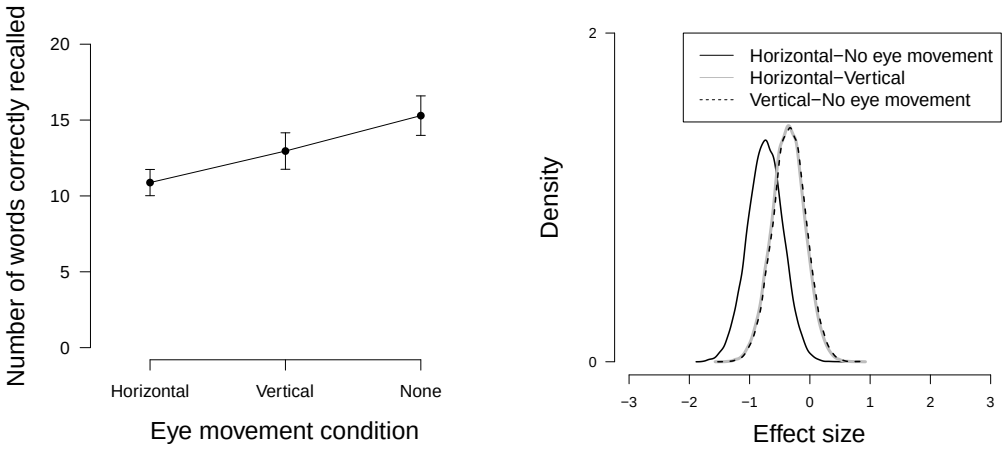
Figure Captions

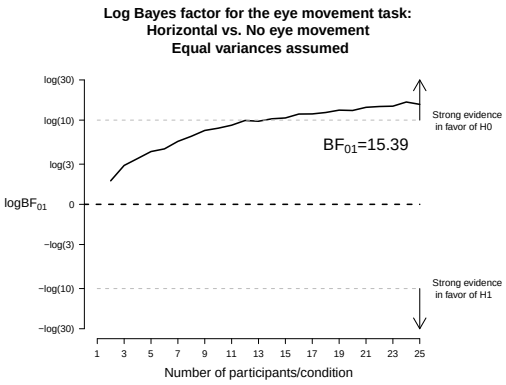
Figure 1. Mean number of words recalled correctly and effect sizes in the three eye movement conditions. The left panel shows the average number of experimental words recalled correctly in the three eye movement conditions. The error bars indicate the standard error. The right panel shows the posterior distribution of the estimated effect size for the horizontal−no eye movement comparison (solid black line), for the horizontal−vertical eye movement comparison (solid gray line), and for the vertical−no eye movement comparison (dashed line).

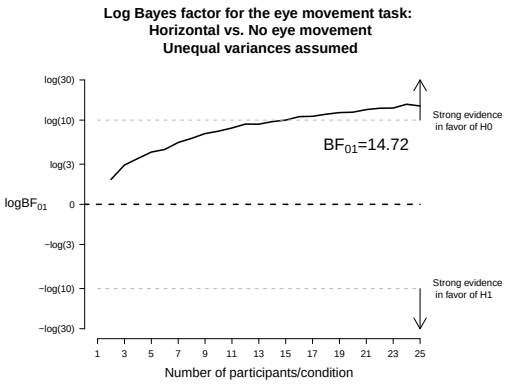
Figure 2. Log Bayes factors (log BF_{01}) for the comparison of the number of correctly recalled words between the horizontal, vertical, and no eye movement conditions.

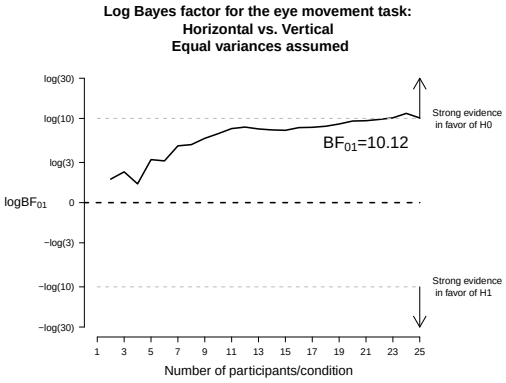
Figure 3. Log Bayes factors (log BF_{01}) as a function of the standard deviation (sd) of the zero-centered normal prior on effect size. Equal variances are assumed. The black triangle corresponds to the Bayes factor computed with a standard normal prior (i.e., unit-information prior) on effect size.

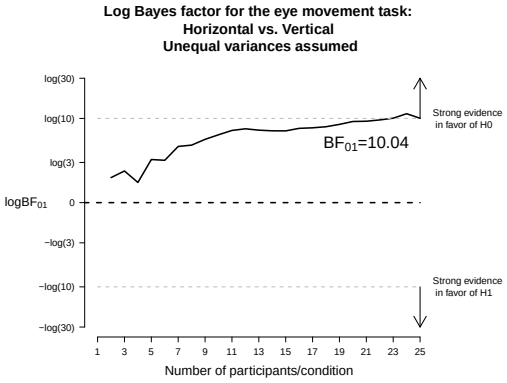
Figure 4. P-curve: The distribution of statistically significant p values in the eye movement literature. The p-curve shows the percentage of significant p values on the intervals $p < .01$, $.01 \leq p < .02$, $.02 \leq p < .03$, $.03 \leq p < .04$, $.04 \leq p < .05$. The exact p values in a given interval are printed above the corresponding percentage.

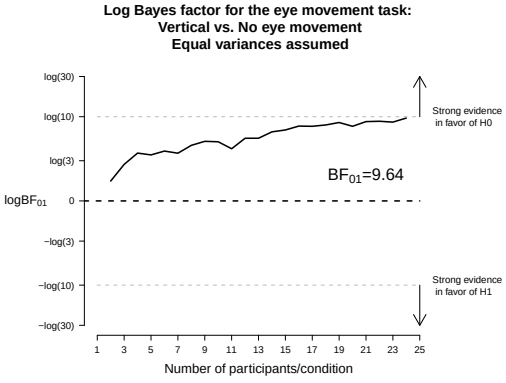

Number of words correctly recalled
20
15
10
5
0
Horizontal
Vertical
None
Eye movement condition
Density
2
0
Horizontal−No eye movement
Horizontal−Vertical
Vertical−No eye movement
−3
−2
−1
0
1
2
3
Effect size

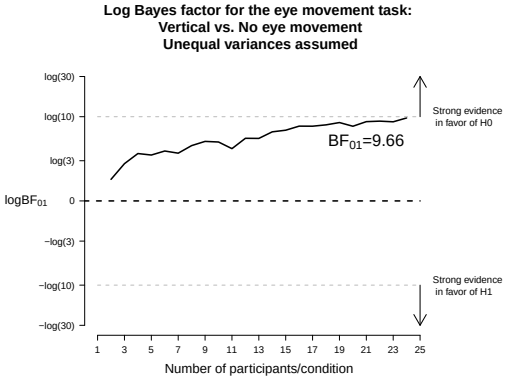

Log Bayes factor for the eye movement task:
Horizontal vs. No eye movement
Equal variances assumed
log(30)
log(10)
log(3)
logBF01
0
−log(3)
−log(10)
−log(30)
Strong evidence
in favor of H0
Strong evidence
in favor of H1
BF01=15.39
1 3 5 7 9 11 13 15 17 19 21 23 25
Number of participants/condition

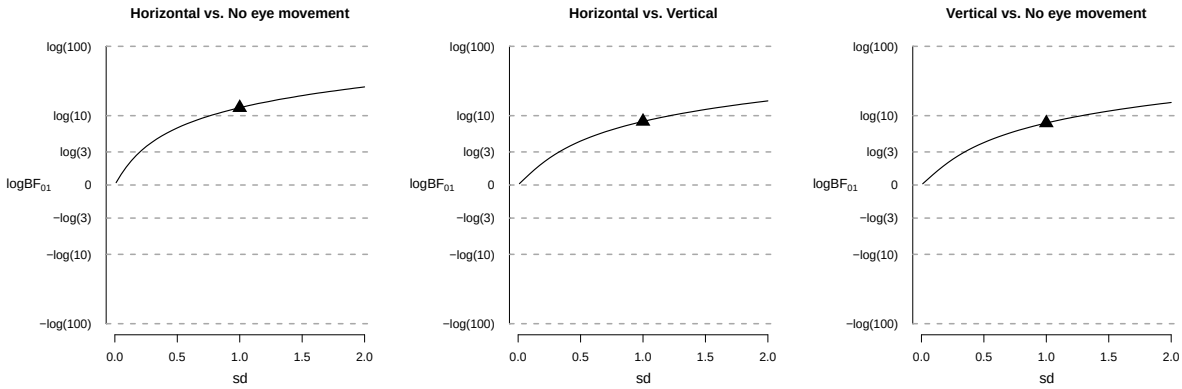

Log Bayes factor for the eye movement task:
Horizontal vs. No eye movement
Unequal variances assumed
log(30)
log(10)
log(3)
logBF01
0
−log(3)
−log(10)
−log(30)
Strong evidence
in favor of H0
Strong evidence
in favor of H1
BF01=14.72
1 3 5 7 9 11 13 15 17 19 21 23 25
Number of participants/condition

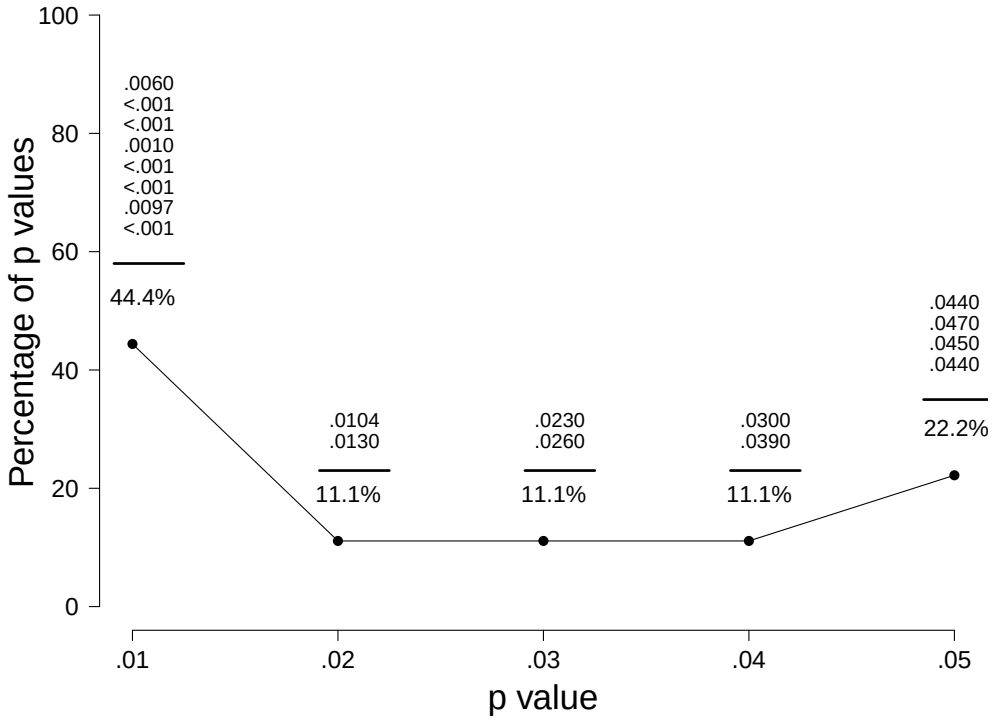

Log Bayes factor for the eye movement task:
Horizontal vs. Vertical
Equal variances assumed
log(30)
log(10)
log(3)
logBF01
0
−log(3)
−log(10)
−log(30)
Strong evidence
in favor of H0
Strong evidence
in favor of H1
BF01=10.12
1 3 5 7 9 11 13 15 17 19 21 23 25
Number of participants/condition


Log Bayes factor for the eye movement task:
Horizontal vs. Vertical
Unequal variances assumed
log(30)
log(10)
log(3)
logBF01
0
−log(3)
−log(10)
−log(30)
Strong evidence
in favor of H0
Strong evidence
in favor of H1
BF01=10.04
1 3 5 7 9 11 13 15 17 19 21 23 25
Number of participants/condition


Log Bayes factor for the eye movement task:
Vertical vs. No eye movement
Equal variances assumed
log(30)
log(10)
log(3)
logBF01
0
−log(3)
−log(10)
−log(30)
Strong evidence
in favor of H0
Strong evidence
in favor of H1
BF01=9.64
1 3 5 7 9 11 13 15 17 19 21 23 25
Number of participants/condition


Log Bayes factor for the eye movement task:
Vertical vs. No eye movement
Unequal variances assumed
log(30)
log(10)
log(3)
logBF01
0
−log(3)
−log(10)
−log(30)
Strong evidence
in favor of H0
Strong evidence
in favor of H1
BF01=9.66
1 3 5 7 9 11 13 15 17 19 21 23 25
Number of participants/condition


Horizontal vs. No eye movement
Horizontal vs. Vertical
Vertical vs. No eye movement
log(100)
log(10)
log(3)
logBF01
0
−log(3)
−log(10)
−log(100)
0.0
0.5
1.0
1.5
2.0
sd


100
80
60
40
20
0
Percentage of p values
.0060
<.001
<.001
.0010
<.001
<.001
.0097
<.001
44.4%
.0104
.0130
11.1%
.0230
.0260
11.1%
.0300
.0390
11.1%
.0440
.0470
.0450
.0440
22.2%
.01
.02
.03
.04
.05
p value