Bayesian explorations in mathematical psychology

Matzke, D.

Publication date
2014
Document Version
Final published version

Chapter 10

Two Birds with One Stone: A Preregistered Adversarial Collaboration on Horizontal Eye Movements in Free Recall

This chapter has been submitted for publication as:
Dora Matzke, Sander Nieuwenhuis, Hedderik van Rijn, Heleen A. Slagter, Maurits W. van der Molen, and Eric-Jan Wagenmakers (2013).
Two birds with one stone: A preregistered adversarial collaboration on horizontal eye movements in free recall.

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.

10.1 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 et al., 2011).

10.2 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 reply-rejoinder disputes— urged the scientific community 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; Tetlock & Mitchell, 2009; 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, 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, 1961a, 1961b). 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, Munafo, & et al., 2013; de Groot, 1961a; Goldacre, 2009; Ioannidis, 2005; Koole & Lakens, 2012; Nosek, Spies, & Motyl, 2012; Wagenmakers et al., 2011). 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://opensciencframework.org`).

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 for 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.

10.3 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., C. W. Lee & Cuijpers, 2013). During EMDR, the patient is 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 the patient's 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. 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 can be downloaded at `http://opensciencefframework.org/project/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.

Data Analysis

In adversarial collaborations is it essential to be able to quantify evidence in favor of the null hypothesis. Moreover, it is desirable to collect data until the pattern of results is sufficiently clear. Neither requirement can be accomplished within the framework of frequentist statistics. The present experiment therefore relied on Bayesian hypothesis testing using the Bayes factor (e.g., Berger & Mortera, 1999; Edwards et al., 1963; Jeffreys, 1961; Kass & Raftery, 1995; Rouder et al., 2012, 2009; Wagenmakers, 2007; Wagenmakers et al., 2010, 2011, 2012; Wetzels et al., 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).[1] For instance, $BF_{01} = 10$ indicates that the data are 10 times more likely under the null hypothesis than under the alternative hypothesis. Alternatively, $BF_{01} = \frac{1}{10}$ indicates that the data are 10 times more likely under the alternative hypothesis than under the null hypothesis. Within the framework of Bayesian inference, the intention with which the data are collected is irrelevant (Edwards et al., 1963); hence we can monitor the Bayes factor as the data are collected (i.e., sequential hypothesis testing), and may stop testing whenever the evidence is sufficiently compelling.

Accordingly, the adversaries 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 the null or the alternative hypothesis (see Jeffreys, 1961, for a classification scheme for the Bayes factor). Specifically, the two parties agreed to stop data collection whenever $BF_{01} > 10$ or $BF_{01} < .1$ for the horizontal vs. no eye movement condition comparison. The adversarial collaboration agreement contains the precise specification of the stopping rule.

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 et al., 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 the alternative hypothesis 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). The adversaries tested the presence of the associative-priming effect using a one-sided paired-sample Bayesian t test as specified by Wetzels et al. (2009).

10.4 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").

[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.

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

Free Recall and Eye Movement Task

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).[2] 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.

Associative-Priming Task

The stimulus pool consisted of 64 prime-word pairs and 64 prime-nonword pairs. The primes and the word targets were Dutch nouns, while the nonwords were pseudowords derived from Dutch nouns by changing one or two letters. The nonwords were generated using the Wuggy software (Keuleers & Brysbaert, 2010). In all prime-word pairs, the target word appeared as an associate of the prime in the Dutch word association norms (de Groot, 1980). The prime-word pairs were adopted from de Groot (1984, 1987). The primes for the prime-nonword pairs were unrelated to the prime-word pairs and to the nouns that were used to create the nonwords. One subset of the prime-nonword pairs was adopted from de Groot (1984), whereas the other subset was selected uniquely for the purpose of the present experiment.

The stimulus pool was used to create two lists that each contained 32 related prime-word pairs and 32 unrelated prime-word pairs. The unrelated word pairs were created by rearranging the primes and the word targets of 32 of the 64 related prime-word pairs. Each target word thus appeared in both lists, either as a target in a related prime-word pair or as a target in an unrelated prime-word pair. The length and frequency of the target words were equated across the related

[2]The stimulus words are available from the adversarial collaboration agreement.

and unrelated prime-word pairs in both lists. The associative strength of the related prime-word pairs was equated across the two lists. The same prime-nonword pairs were used across the two lists. Word length was equated across nonwords and the target words in the prime-word pairs.[3]

The two stimulus lists were counterbalanced across participants. The prime-word pairs and the prime-nonword pairs were presented sequentially on a computer screen. The order of stimulus presentation was randomized across participants. The stimuli were presented in black using lowercase 34 point Arial in the middle of a light-gray display. First, a fixation cross appeared on the screen for 1000 ms slightly above and left of the position of the to-be-presented prime, followed by a blank inter-stimulus interval of 20 ms. Next, the prime appeared in the middle of the screen for 400 ms, followed by a blank inter-stimulus interval of 40 ms. Next, the target appeared slightly below the position of the previously presented prime. The target remained on screen until the participant responded or until 2400 ms elapsed. Participants were instructed to press 'M' with their right index finger for 'word' responses and to press 'Z' with their left index finger for 'nonword' responses. Incorrect responses were followed by the message 'FOUT' (i.e., incorrect), responses slower than 1200 ms were followed by the message 'TE LANGZAAM' (i.e., too slow), and responses faster than 200 ms were followed by the message 'TE SNEL' (i.e., too fast). If the participant failed to respond within 2400 ms, 'TE LANGZAAM' appeared automatically on the screen and an error was recorded. The feedback was presented 20 ms after response/target offset, slightly below the position of the previous target. The feedback remained on the screen for 1200 ms. The feedback scheme was intended to promote accurate but fast responding. Following 1000 ms after a correct response or after the offset of an error message, the fixation cross reappeared on the screen.

The experimental stimuli were presented in four blocks of 32 prime-target pairs. A forced rest of 30 sec. separated the experimental blocks. The presentation of the 128 experimental prime-target pairs was preceded by a practice list of 32 prime-target pairs. The practice list consisted of 8 related prime-word pairs, 8 unrelated prime-word pairs, and 16 prime-nonword pairs, none of which were also present in the 128 experimental word-target pairs.

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. For the eye movement task, participants were instructed to follow the circle with their eyes by making saccadic eye movements and to keep their head still. The experimenter carefully monitored participants' compliance with the instructions, including the eye movement behavior.

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. Instructions emphasized fast but accurate responding. Participants were instructed to pay attention to both letter strings (i.e., prime and target), but only respond to the second letter string (i.e., the target). The instructions did not mention the association between the related prime-word pairs. Lastly, participants completed an exit interview, inquiring about their age and gender. In addition, par-

[3]The associative-priming stimuli are available from the adversarial collaboration agreement.

ticipants 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.

10.5 Results

Confirmatory Analyses

Eye Movement Task

The free recall data are available on the OSF at `http://opensciencefframework.org/project/pXT3M/`. Based on the exclusion criteria specified in the adversarial collaboration agreement, we excluded the free recall data of two participants who correctly described the goal of the experiment and four participants who recalled fewer than five items correctly. The analyses reported below are based on the data of 25 participants in the horizontal eye movement ($N_H = 25$), 24 participants in the vertical eye movement ($N_V = 24$), and 24 participants in the no eye movement condition ($N_F = 24$).

The left panel of Figure 10.1 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 (4.32) in the horizontal, 12.96 (5.89) in the vertical, and 15.29 (6.38) in the no eye movement condition. The right panel of Figure 10.1 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. Figure 10.2 shows 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 the alternative hypothesis, whereas log Bayes factors higher than zero indicate evidence for the null hypothesis.

For all three hypotheses, the evidence in favor of the null hypothesis gradually increased as the data accumulated. After testing 73 participants, the Bayes factor indicated that the data are 15 times more likely under the null hypothesis of no difference between the horizontal and the no eye movement condition than under the alternative hypothesis ($BF_{01} = 15.39$).[4] Similarly, the Bayes factor indicated that the data are more than 10 times more likely under the null hypothesis of no difference between the horizontal and the vertical eye movement condition than under the alternative hypothesis ($BF_{01} = 10.12$). Lastly, the Bayes factor indicated that the data are more

[4]After five weeks of data collection, the 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://opensciencefframework.org/project/pXT3M/`

than 9 times more likely under the null hypothesis of no difference between the vertical and the no eye movement condition than under the alternative hypothesis ($BF_{01} = 9.64$). As shown in the right panels of Figure 10.2, essentially the same results were obtained under the assumption of unequal variances. Unsurprisingly, the frequentist alternatives of the one-sided unpaired Bayesian t 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 the null hypothesis 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 the null hypothesis for the vertical vs. no eye movement comparisons.

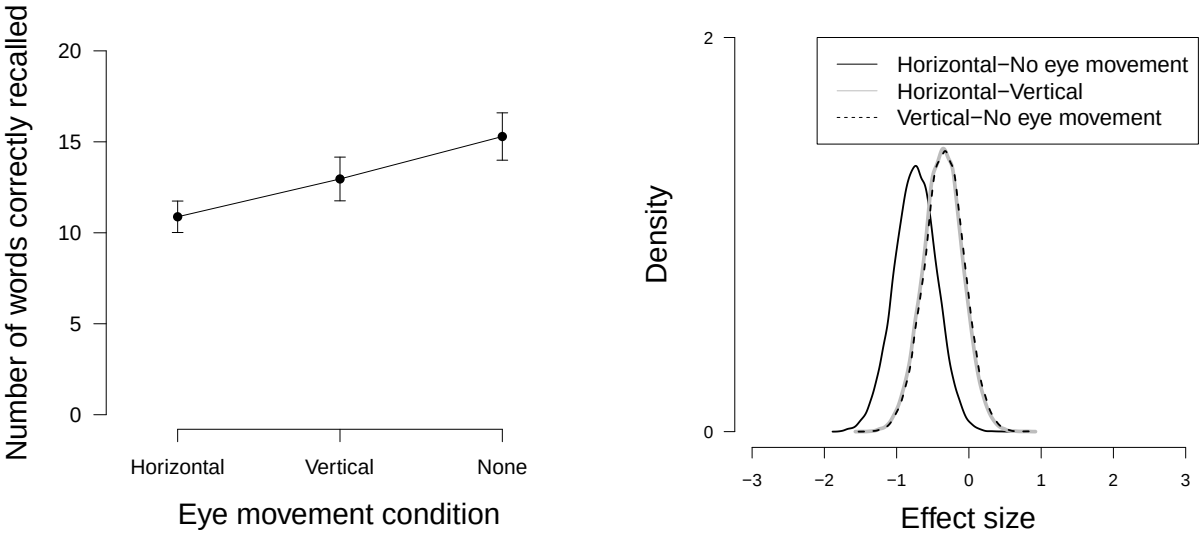

Figure 10.1 *Mean number of words recalled correctly and effect sizes in the three eye movement conditions.* The left panel shows the average number of experimental words recalled correctly in the three eye movement conditions. The error bars indicate the standard error. The right panel shows the posterior distribution of the estimated effect size for the horizontal−no eye movement comparison (solid black line), for the horizontal−vertical eye movement comparison (solid grey line), and for the vertical−no eye movement comparison (dashed line).

Associative-Priming Task

The priming data are available on the OSF at `http://opensciencframework.org/project/pXT3M/`. We used only correct RTs that were longer than 250 ms and shorter than 1500 ms, resulting in an average exclusion rate of 6.39%. Based on the exclusion criteria specified in the adversarial collaboration agreement, we excluded one participant with error rate higher than 20%. We excluded the data of one additional participant because of computer failure. The analysis reported below is based on 77 participants.

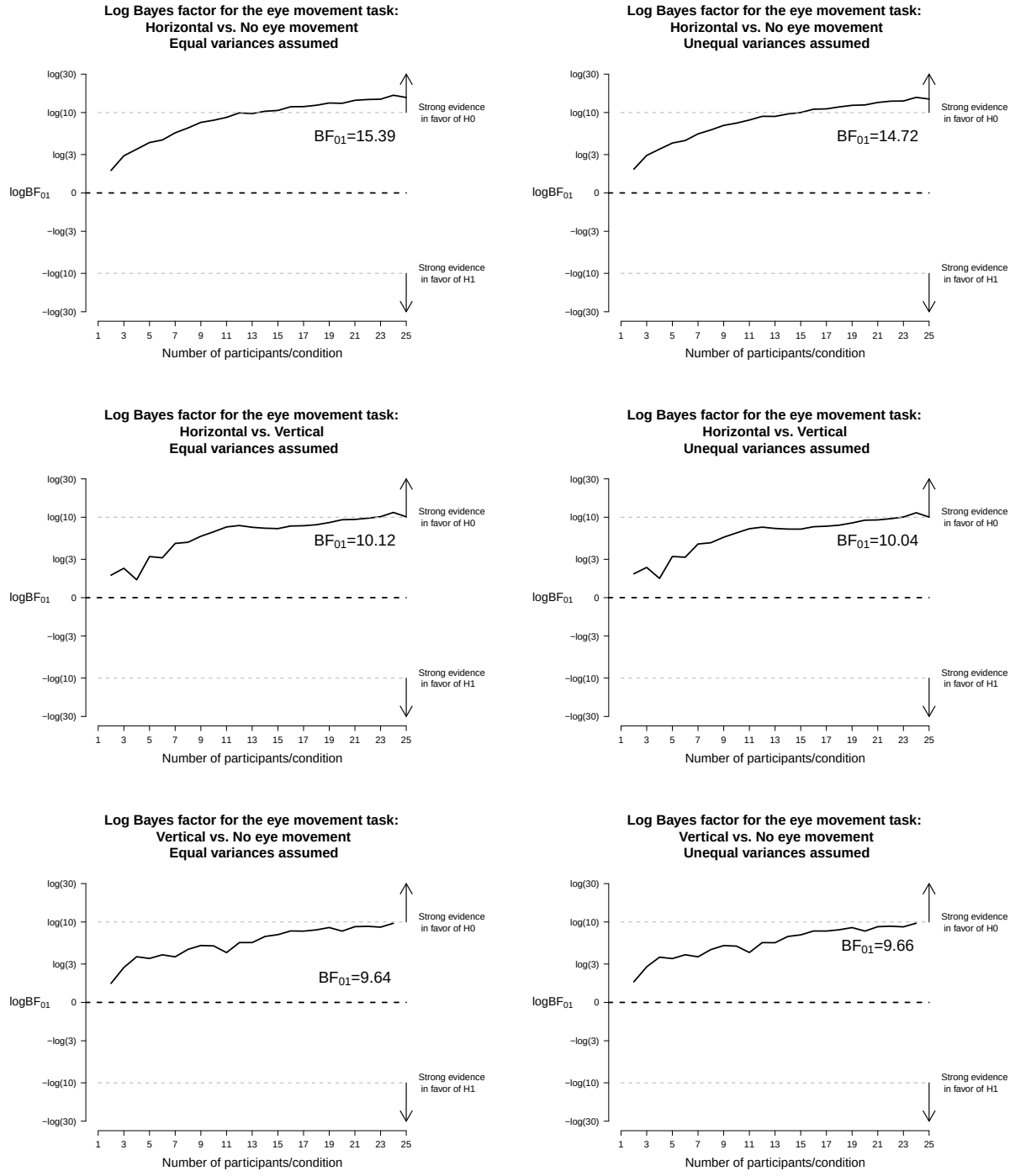

Figure 10.2 *Log Bayes factors for the comparison of the number of correctly recalled words between the horizontal, vertical, and no eye movement conditions.*

Figure 10.3 shows mean RT for the related and the unrelated prime-word pairs and the corresponding effect size. As expected, 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,

sd = 66.35). The posterior distribution assigns most mass to large negative effect sizes. Figure 10.4 shows the results of the sequential analysis using Bayes factors from the default one-sided paired-sample Bayesian t test. As the data accumulated, the evidence for the alternative hypothesis gradually increased. After testing 77 participants, the Bayes factor indicated that the data are 528,848,417 times more likely under the alternative hypothesis than under the null hypothesis ($BF_{01} = 1.890901E-09$). This result supports the adversaries' expectation and indicates extreme evidence for the presence of the associative-priming effect.

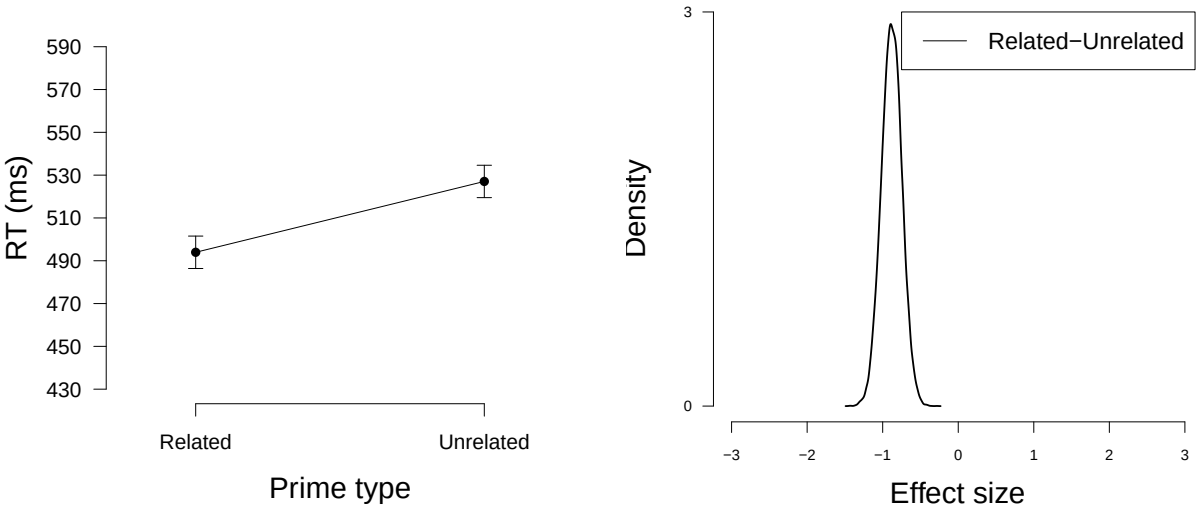

Figure 10.3 *Mean RT and effect size for the associative-priming task.* The left panel shows mean RT for the related and the unrelated prime-word pairs. The error bars indicate the standard error. The right panel shows the posterior distribution of the estimated effect size for the related−unrelated prime-word comparison.

Exploratory Analyses

This section presents 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 eye movement 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 shape 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 the null hypothesis. Thus, highly uninformative prior distributions yield Bayes factors that lend infinite support for the null hypothesis (Jeffreys, 1961).

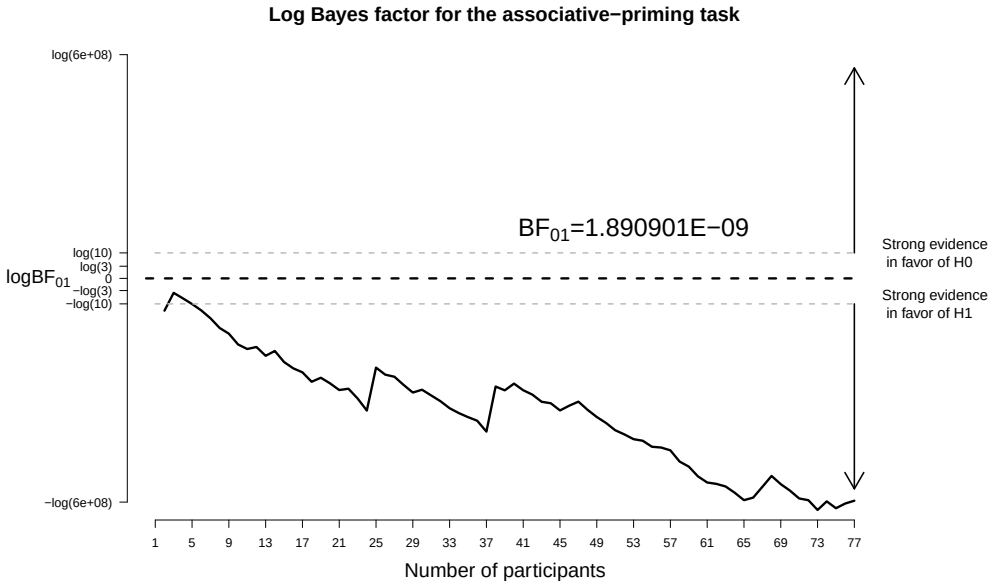

Figure 10.4 *Log Bayes factor for the comparison of mean RT for related vs. unrelated prime-word pairs.*

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., $\delta \sim \text{Normal}(0, \text{sd})^+$).

Figure 10.5 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 the alternative hypothesis, whereas log Bayes factors higher than zero indicate evidence for the null hypothesis. 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 the null hypothesis. 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 the alternative hypothesis over the null hypothesis regardless of the variability of the prior distribution. Even under the prior setting that maximally supports the alternative hypothesis (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 the null hypothesis varied as a function of the prior setting, the Bayes factor always favored the null hypothesis over the alternative hypothesis regardless of the variability of the prior.

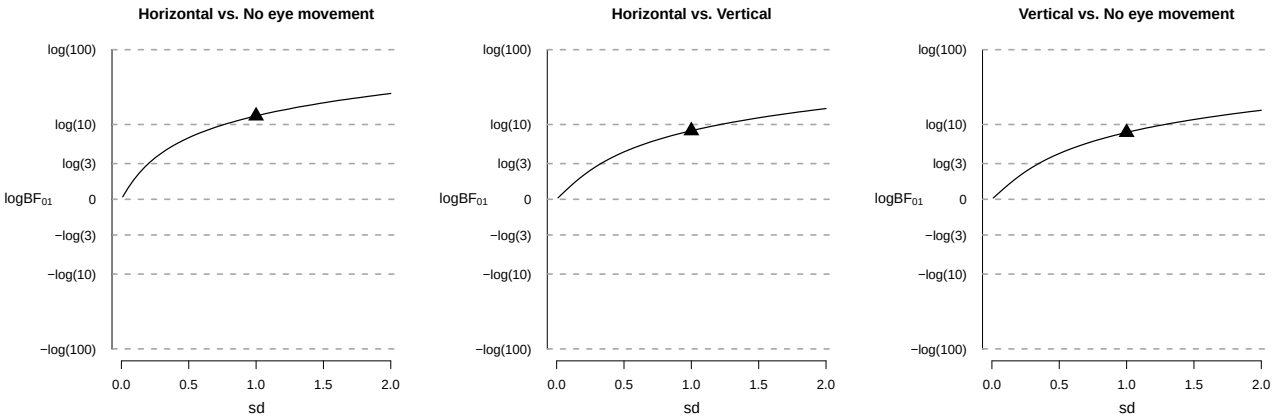

Figure 10.5 *Log Bayes factors (log BF_{01}) as a function of the standard deviation (sd) of the zero-centered normal prior on effect size.* Equal variances are assumed. The black triangle corresponds to the Bayes factor computed with a standard normal prior (i.e., unit-information prior) on effect size.

10.6 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 the null hypothesis 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 the null hypothesis for the vertical eye movement vs.

no eye movement comparison. Lastly, the results of the associative-priming task supported the adversaries' 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. Our joint study thus failed to replicate the beneficial effect of bilateral eye movements on episodic memory. Despite our best efforts to carry out a prototypical experiment, the present study —and our two pilot studies— contradicts the seemingly well-established finding on the association between horizontal eye movements and memory retrieval.

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, 201s). 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 the null hypothesis (Berger & Delampady, 1987; Edwards et al., 1963; V. E. Johnson, 2013; Sellke et al., 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 Bayes factors that indicate that the data are no more than three times more likely under the alternative hypothesis than under the null hypothesis. 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). 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). Error mechanisms during the interpretation of the data, such as hindsight bias and positive confirmation bias, 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), and the use of post-hoc exclusion criteria, such as arbitrary handedness cut-off scores.[5]

[5]The following investigations all used different criteria for classifying participants as strongly right-handed:

The present paper therefore emphasizes the importance of preregistration and the strict separation of confirmatory and exploratory research (see also de Groot, 1961a).

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 current 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 effect 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 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 5). 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.

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.

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 as supplemental material at `http://dora.erbe-matzke.com/publications.html`. Figure 10.6 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,

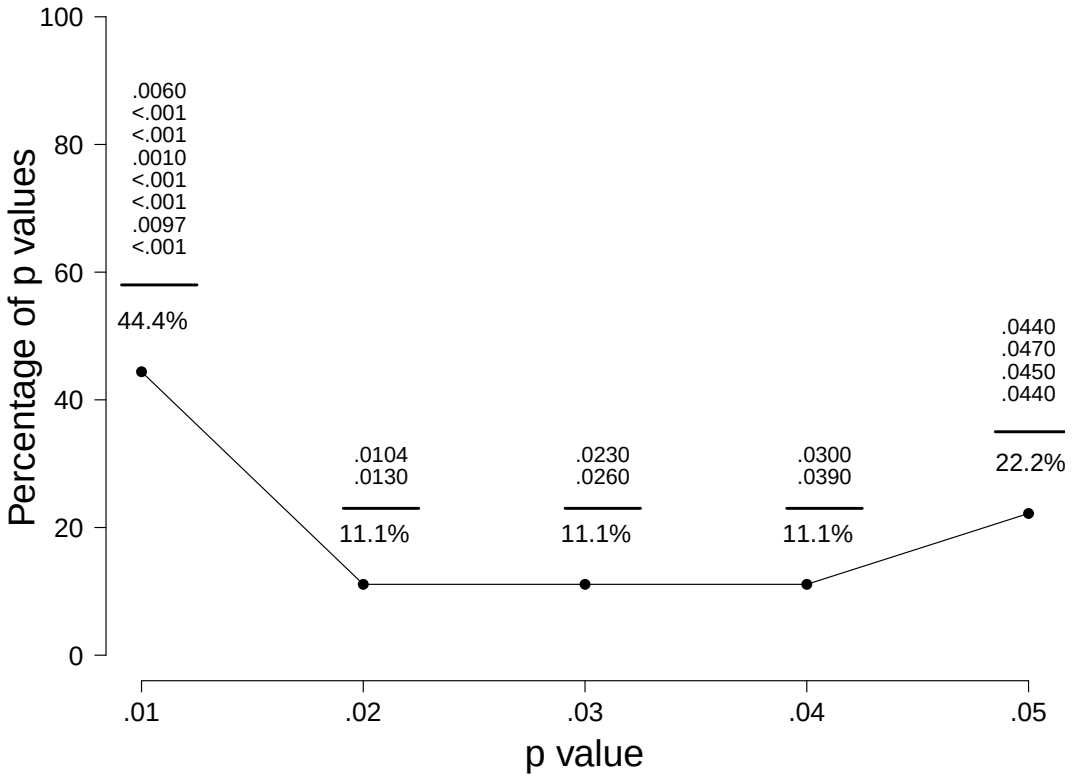

Figure 10.6 *P-curve: The distribution of statistically significant p values in the eye movement literature.* The *p*-curve shows the percentage of significant p values on the intervals $p < .01$, $.01 \leq p < .02$, $.02 \leq p < .03$, $.03 \leq p < .04$, $.04 \leq p < .05$. The exact p values in a given interval are printed above the corresponding percentage.

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; C. W. 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. In retrospect, then, we have to conclude that the adversarial collaboration agreement was not watertight. It should have specified the conditions under which the parties would have been prepared to give up their point of view. If a single failure to replicate, based upon a strict agreement concerning the particulars of the experiment and associated data analysis, is not sufficient, the obvious danger is to encounter a situation well described by an unknown quote "Theories are like old soldiers, they never die but slowly fade away".