The prioritization of visuo-spatial associations during mental imagery

Hafidah Umar[1,2,3], Fred W. Mast[3], Trix Cacchione[4], & Corinna S. Martarelli[5]

[2] Department of Neurosciences, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian, Kelantan, Malaysia

[2] Brain and Behaviour Cluster, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian, Kelantan, Malaysia

[3] Department of Psychology, University of Bern, Bern, Switzerland

[4] Department of Developmental Psychology, School of Education, University of Applied Sciences and Arts Northwestern Switzerland, Windisch, Switzerland

[5] Faculty of Psychology, Swiss Distance University Institute, Brig, Switzerland

Correspondence: Corinna Martarelli, Swiss Distance University Institute, Überlandstrasse 12, 3900 Brig, Switzerland, Phone: +41 (0)76 340 09 82, Email: corinna.martarelli@fernuni.ch

Umar, H., Mast, F. W., Cacchione, T., & Martarelli, C. S. (2021). The prioritization of visuo-spatial associations during mental imagery. *Cognitive Processing.* https://doi.org/10.1007/s10339-020-01010-5

The prioritization of visuo-spatial associations during mental imagery

Abstract

While previous research has shown that during mental imagery participants look back to areas visited during encoding it is unclear what happens when information presented during encoding is incongruent. To investigate this research question, we presented 30 participants with incongruent audio-visual associations (e.g. the image of a car paired with the sound of a cat) and later asked them to create a congruent mental representation based on the auditory cue (e.g. to create a mental representation of a cat while hearing the sound of a cat). The results revealed that participants spent more time in the areas where they previously saw the object and that incongruent audio-visual information during encoding did not appear to interfere with the generation and maintenance of mental images. This finding suggests that eye movements can be flexibly employed during mental imagery depending on the demands of the task.

Keywords: visual mental imagery, audio-visual integration, eye movements, corresponding area effect, looking back to nothing effect.

Introduction

Mental imagery has been defined as a conscious sensory experience, in which mental images are reconstructed in the absence of actual corresponding sensory stimulations (Lacey & Lawson, 2013). Imagery activity is considered to rely on similar motor processes (i.e. eye movements) to the ones used during perception, with oculomotor experiences happening gradually and sequentially during both imagery and perceptual processes alike (Brandt & Stark, 1997; Laeng, Bloem, D'Ascenzo, & Tommasi, 2014). Hebb (1968), for instance, pointed out that when forming a mental image of a familiar object (such as a car), the internal representation of that image has to be sequentially constructed and organized. In fact, similar processes are at work when we look at objects during the process of perception and when we make a series of fixations toward different parts of an object. Thus, the scanning process relies on mechanisms that are partly shared by visual perception and mental imagery (Borst & Kosslyn, 2008). Eye fixations fall sequentially from one part to the other of a mental object, producing a series of eye movements (Laeng & Teodorescu, 2002).

Richardson and Spivey (2000) focused on eye movements during recall of auditory information (i.e. factual sentences). In line with their work, Scholz, Klichowicz, and Krems (2018) (see also Scholz, Mehlborn, & Krems, 2016; see Kumcu & Thompson, 2016, 2018 for a different approach with verbal material) found that auditory information is locked to visuo-spatial information in mental representations. In their study participants listened to sentences that were associated with one of four spatial locations on a computer screen with a loudspeaker symbol (visual cue). In the retrieval phase, participants had to judge whether statements about the previously heard sentences were true or false. Even though there was no semantic relation between the presented auditory statement and the visual cue, participants showed the typical looking back to nothing phenomenon (i.e. the eyes revisited the locations where the visual cues were presented, when the sentences were later retrieved). This suggests that auditory information is encoded with a spatial memory trace that can be used for later retrieval of information. Spatial information has been conceptualized as a deictic pointing device that is *automatically* associated with the memory trace during encoding (e.g. Ballard, Hayhoe, Pook, & Rao, 1997). When the event is absent, the deictic pointing device triggers eye movements to the spatial location of the previously encoded event. The function of deictic pointing devices is to bind objects and sounds in the world to cognitive processes such as memory retrieval. Executing eye movements to the areas where the stimuli were previously encoded activates parts of a multisensory memory trace and thus might increase memory performance (e.g. Ferreira, Apel, Henderson, 2008; Richardson, Altmann, Spivey, & Hoover, 2009).

The present study aims to investigate the use of multiple representations and to investigate prioritization of information during mental imagery. It is still unclear whether incongruent information during encoding will affect people's tendency to look back to nothing during mental imagery. For investigating this research question, we used pairs of incongruent stimuli (e.g. the image of a cat that was paired with the sound of a car, and the image of a car that was paired with the sound of a cat) in which the image appeared at different locations on the screen. Participants were then instructed to generate mental images corresponding to the heard sounds and we measured whether the triggering sounds prompt looking back to the areas in which the images had been previously seen vs. to the areas where other incongruent images were located when the same sound was presented during encoding. In their recent work, Ryan and Shen (2020) highlight the advantages of measuring gaze behavior to investigate memory processes, in particular because gaze behavior conveys information about the contents of visual memories and can thus reveal multiple ongoing representations that differ from the final task decision. If the auditory information is encoded with the spatial memory trace (e.g. Ferreira et al., 2008), we can expect an automatic looking back to the areas associated with the sound, at least during the first fixations. In addition to recording eye movements, we assessed the accuracy of the mental representations by asking detailed questions about the previously seen stimuli (objective measure of mental imagery). Moreover, we assessed the vividness of the mental representations via subjective judgements (subjective measure of mental imagery).

Given the nature of our task, we hypothesize that spatial indexing by eye movements will mainly be guided by visual information and thus we expect that eye movements during mental imagery will mostly be directed to the areas where the corresponding visual stimuli were previously presented. Based on empirical evidence summarized above (e.g. Scholz et al., 2018), we hypothesize that spatial indexing by eye movements will also be guided by auditory information. Thus, we expect that participants will show eye movements to areas that were previously associated with specific sounds, when compared to non-corresponding locations. More generally, we expect the looking behavior during mental imagery to mirror this competition between two distinct spatial locations.

Method

Participants

Thirty individuals (24 females, 6 males) participated in the study (mean age = 21.40, *SD* = 1.79). All participants gave informed consent according to the guidelines of the University of Bern institutional ethics review board. All participants reported normal or corrected-to-normal vision and hearing. None reported a color vision deficiency. They were all naïve regarding the purpose of the study. Participants received a course credit in return for their participation. The experimental procedure took approximately 90 minutes to complete. The study was approved by the university's local ethics committee and was carried out in accordance with the Declaration of Helsinki of 1975.

Apparatus

Eye data were recorded using a video-based iView X RED tracking system (SensoMotoric Instruments, Teltow, Germany), which was integrated with a 22-inch TFT monitor (1680 × 1050 pixels) for the presentation of visual stimuli. The system captured the eye data with a sampling rate of 250 Hz, a tracking resolution of < 0.01º, and a gaze position accuracy of < 0.5º.

Stimulus Materials

During the *Encoding phase*, participants were presented with audio-visual stimuli. To create the stimuli, we firstly selected 32 images of animals and 32 images of non-living objects (i.e. vehicles, musical instruments, and household items) from the internet. All images with a scene background were edited to make the background transparent, using Inkscape software. Then, each image was cropped and saved in a PNG format. Secondly, we used PowerPoint (Microsoft Office) to prepare a blank white workspace (34 cm width × 27 cm height). In order to determine the location where the individual image had to be placed, the visual field was divided into four identical quadrants along the vertical and horizontal midlines (i.e. upper-left, upper-right, bottom-left, and bottom-right). Each of the animal images were resized proportionately and positioned at the center of one of the quadrants (see Figure 1, for an example of the stimuli).

The auditory stimuli were sounds usually associated with the 32 animals and 32 non-living objects. The sounds were obtained from different internet sources, edited and saved as MPEG Layer 3, stereo format at a sampling rate of 44,100 Hz. Each sound was 5000 ms in duration. The selection of the animal images and sounds mostly followed Viggiano et al. (2017), except for goose, rooster and turkey (i.e. 9.38 % of the animal stimuli used). Similarly, the selection of the images and sounds of non-living objects was taken from

Viggiano et al. (2017), except for bicycle, fire truck, ship, bagpipes, drum, guitar, tambourine, alarm clock, lawn mower and printer (i.e. 31.25 % of the non-living object stimuli used). Slight changes in the selection of the animal images and sounds when compared to Viggiano and colleagues were due to differences in the aims of the studies.

We then performed two types of incongruent audio-visual manipulations: Intra-categorical manipulations (e.g. the image of a tiger paired with the sound of a horse) and extra-categorical (e.g. the image of a cat paired with the sound of a car). Half of the bimodal stimuli were intra-categorical (32 stimuli) and half were extra-categorical (32 stimuli). To avoid the impact of idiosyncrasies of a reduced set of stimulus material, we created two sets of 32 bimodal stimuli. To control for the impact of intra-categorical and extra-categorical manipulations we created the intra-categorical (Group A) and extra-categorical (Group B) conditions. Participants were randomly assigned to either Group A or Group B (between-subject design, n = 15 in each condition). We selected animals and objects to be included in Group A and Group B considering the physical characteristics and sounds/vocalizations produced by each animal or object, avoiding the use of similar animals or objects in the same group (e.g. if Group A contained a tiger and a goat, the lion and the sheep were assigned to Group B).

In both groups two variants of image-sound pairs were created. Both variants included the same two stimuli (e.g. cat and car), but they were presented in different modalities (as image or sound), and images were located at different quadrants (the screen was divided into four equally sized quadrants). For example, an image-sound pair of ‘cat-car’ was presented as a cat image paired with a car engine sound (with the cat image being located in e.g. the upper-left quadrant), and as a car image paired with a cat sound (with the car image being located in e.g. the bottom-right quadrant). The position of the images belonging to the same pair was never at the same place but in another quadrant out of three. The other quadrant was selected with a counterbalancing technique, so that the image could appear on the opposing quadrant or in an adjacent quadrant in a balanced manner. Please find in Appendix A, a list of the stimuli used during the experiment as well as the position of the images on the screen.

During the *Imagery phase*, participants were required to complete three tasks: The Image Generation task, the Image Inspection task, and the Vividness Rating task. For each of these tasks, participants were presented with a blank white screen, and only heard the sound of the stimuli during the Image Generation task. All sounds presented during encoding were used during the Imagery phase.

Procedure

Participants were tested individually in a dimly lit and quiet sound-attenuated eye-tracking laboratory. They sat at a distance of 70 cm from the screen. The experimental session was preceded by a calibration procedure for the eye-tracker using a 5-point calibration and validation method during which participants had to fixate alternatively at the black calibration dots. Only error values below 0.8° were accepted. Auditory stimuli were presented over closed-ear headphones.

The procedure was divided into a practice phase and a test phase. In the practice phase, participants were exposed to five incongruent audio-visual stimuli to ensure that they were familiar with the eye-tracking experimental setting and that they knew what to expect during the *Encoding phase*. Additionally, the practice phase allowed us to ensure that the decibel level of the sounds (presented via headphones) was appropriate for each participant. The stimuli presented in the practice phase were not presented again in the test phase.

During the *Encoding phase*, 32 audio-visual stimuli were presented for 5000 ms, preceded by a fixation cross presented for 1000 ms at the center of the screen. The order of presentation was fully randomized. Participants were instructed to memorize the images and sounds as precisely as possible.

As mentioned above, participants during the *Imagery phase* were required to complete three tasks: the Image Generation task, the Image Inspection task, and the Vividness Rating task. Overall, 32 trials were administered in this phase. The order of presentation was fully randomized. No time limit was set for the trial duration of these tasks.

Image Generation Task. In the Image Generation task, participants were presented with a blank white screen, which was accompanied by the sound of animals or objects they had previously heard in the Encoding phase. The blank screen was preceded by a fixation cross presented for 1000 ms at the center of the screen. Participants were instructed to generate a mental image on the blank screen. For example, when they heard a 'Meeow', they were expected to visualize the physical features and characteristics of the cat they had seen during the Encoding phase. Participants were also instructed to maintain the mental image in the subsequent two tasks (see below). When they had finished visualizing a particular animal or object, they said 'OK', and the experimenter instantly pressed the space bar to proceed to the next screen.

Image Inspection Task. In the Image Inspection task, participants maintained their mental image on the blank white screen, while hearing a short statement. Participants were instructed to inspect the mental image and to verbally respond to short statements (e.g. "The cat is standing on its two legs") with a 'Yes' or 'No'. After participants had given their response, the next screen was presented.

Vividness Rating Task. In the Vividness Rating task, participants heard the following statement: "How do you rate the vividness of your mental image? Rate from 1 to 5. One is no visualization at all and five is perfectly clear". After participants had given their response, the experiment continued.

Figure 1 is an illustration of the trial sequence. The experimenter recorded participants' response manually. Each task in the Imagery phase was self-paced (i.e. there were no restrictions on speed and duration to accomplish the task). At the end of each trial, a final 'Stop' screen informed participants that they no longer needed to imagine this particular stimulus. On the 'Stop' screen, participants were allowed to blink or rest their eyes before the next trial began.

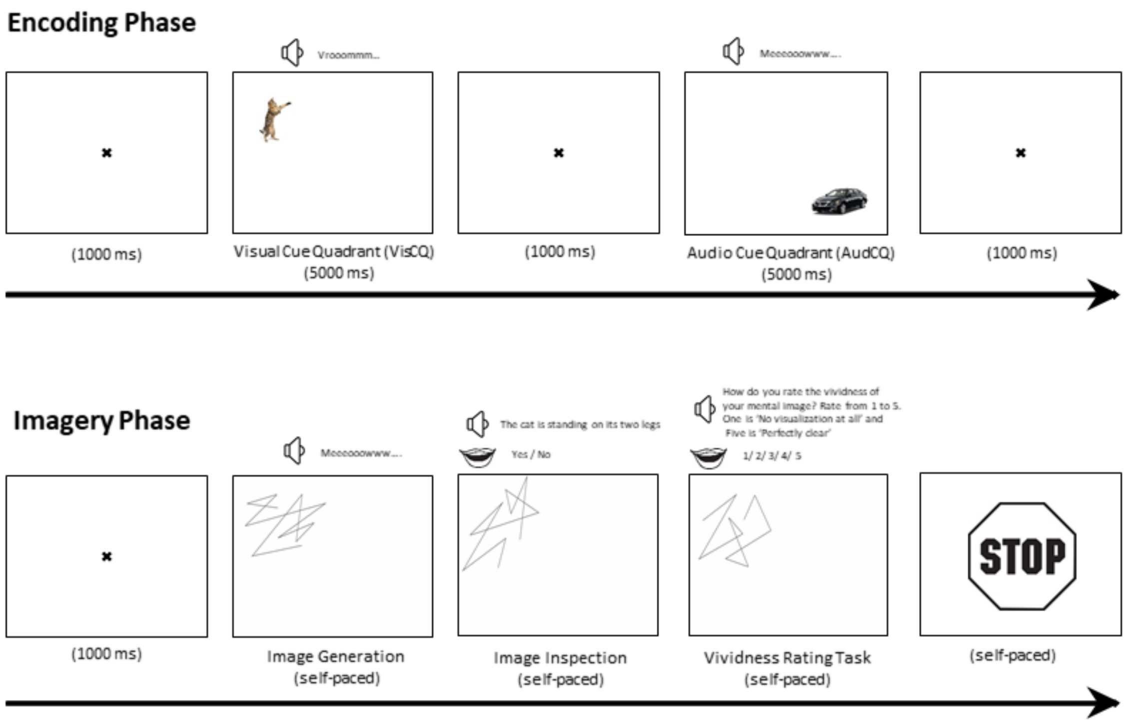

Figure 1. Illustration of the trial sequence in both experimental phases. Encoding phase: the figure shows an example of a 'cat-car' pair, with the image-sound manipulation (extra-categorial group). Note that the presentation of stimuli pairs was fully randomized. Imagery phase: after the fixation screen, the participants were presented with a white blank screen together with the sound of a cat. They were required to form a mental image of the previously seen cat and maintain that mental image until they saw a 'Stop' screen. All auditory stimuli of the Encoding phase were used in the Imagery phase and presented in a randomized order.

Results

Encoding phase: Percentage of dwell time during stimulus presentation

To ensure that participants attended to the presented stimuli, we compared dwell time in the quadrants containing the stimulus, to mean dwell time in the other three quadrants. A paired *t*-test showed that participants spent significantly more time in the quadrant where the image stimuli were located (M = 95.92%, SD = 2.66), as compared to the other quadrants (M = 4.08%, SD = 2.66; $t(29) = 94.42$, $p < .001$, $d = 17.24$). Hence, the analysis confirmed that participants attended to the stimuli during the Encoding phase.

Imagery phase: Percentage of dwell time in the three imagery tasks

Participants were correct in 72.19% (SD = 9.58) of the trials in the Image Inspection task. Analyses were based on correct trials only (there were too few incorrect trials for analyzing them separately). The dependent variable was the mean percentage of dwell time (i.e. the time spent with the eyes in the area of interest, AOI). Depending on the type of stimuli encountered during the previous Encoding phase, quadrants were named as (i) Visual Cue Quadrant (hereafter, VisCQ), if the image had been shown in that quadrant; (ii) Audio Cue Quadrant (hereafter, AudCQ), if the incongruent image paired with the sound had been shown in that quadrant; and (iii) No Cue Quadrant 1 and (iv) No Cue Quadrant 2 (hereafter, NoCQ1 and NoCQ2), if no image had been shown in that quadrant (1 and 2 were attributed following the 'clock coded' method by Richardson and Spivey (2000), starting from the VisCQ). For each participant and trial, we calculated the mean percentage of dwell time in each quadrant. This was analyzed with a mixed ANOVA, with Task (Image Generation task, Image Inspection task, Vividness Rating task) and Cue Quadrant (VisualCQ, AudioCQ, NoCQ1, NoCQ2) as within-subjects factors and Group (intra-categorical, extra-categorical) as between-subjects factor. When the assumption of sphericity was violated ($p < .05$), we used the Huynh-Feldt correction to adjust the degrees of freedom. The effect size was reported as Partial Eta Squared. We used Tukey corrections to adjust for multiple comparisons. The alpha level for statistical tests was set at .05. The dataset for the main analysis can be found at https://figshare.com/s/281e71a7e2d288323488.

In line with our prediction the analysis revealed a significant main effect of Cue Quadrant on dwell time ($F(1.98, 55.37) = 7.41$, $p = .001$, $\eta^2_p = .21$). In particular, dwell time was significantly higher in VisualCQ (34.01%, *SD* 1.65) than in the other Cue Quadrants (AudioCQ: 21.64%, *SD* 0.64, $p = .001$; NoCQ1: 20.98%, *SD* 1.26, $p < .001$; NoCQ2: 23.38%, *SD* 1.53, $p = .006$). Contrary to our predictions, we did not find a longer dwell time in the AudioCQ than in the two other totally irrelevant quadrants (NoCQ1 and NoCQ2). Mean percentages of dwell time across tasks and quadrants are illustrated in Figure 2. The analysis further showed a non-significant interaction between Task × Cue Quadrant × Group, ($F(3.83, 107.20) = 1.09$, $p = .363$, $\eta^2_p = .04$) as well as non-significant 2-way interactions, (Task × Cue Quadrant: $F(3.83, 107.20) = 2.04$, $p = .097$, $\eta^2_p = .07$; Cue Quadrant × Group: $F(1.98, 55.37) = 0.27$, $p = .760$, $\eta^2_p = .01$; Task × Group: $F(1, 28) < 1$). Also the main effect of Group ($F(1, 28) < 1$) and Task ($F(1, 28) < 1$) were not significant.

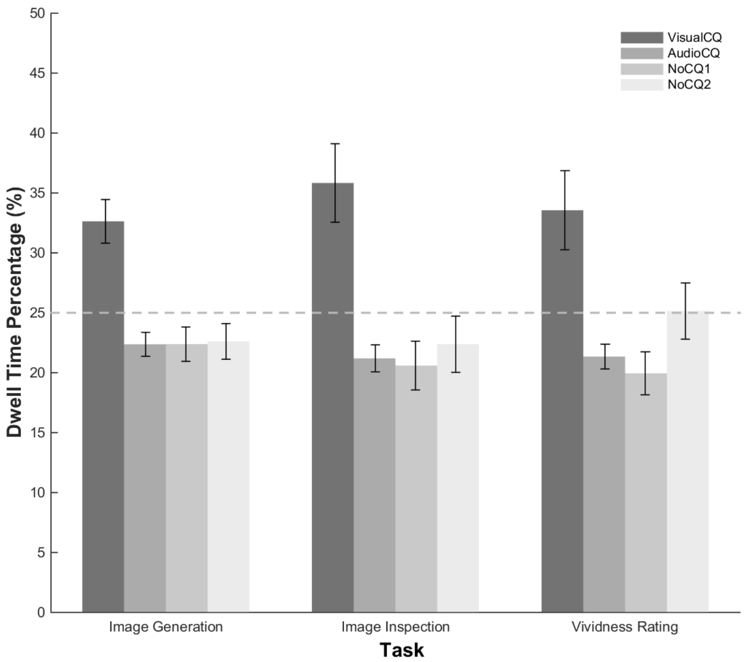

Figure 2. Mean percentage of dwell time in each quadrant during the Image Generation task, the Image Inspection task and the Vividness Rating task in the Imagery phase. Error bars represent ± 1 SEM. The dotted grey line represents chance level.

Contrary to our expectations, we found a similar dwell time in the AudioCQ as in the NoCQ1 and the NoCQ2. To assure that the incongruent visual information during encoding did not play any role in the reconstruction of the mental image, we analyzed the temporal dynamics of the eye pattern on a trial level. It could be possible, for example, that the first fixations are automatically triggered by the incongruent visual information associated with the auditory cue. Figure 3 shows the mean percentage of fixations allocated on the different areas for the first fixation until the fifth fixation separated by task (Image Generation task, Image Inspection task, Vividness Rating task).

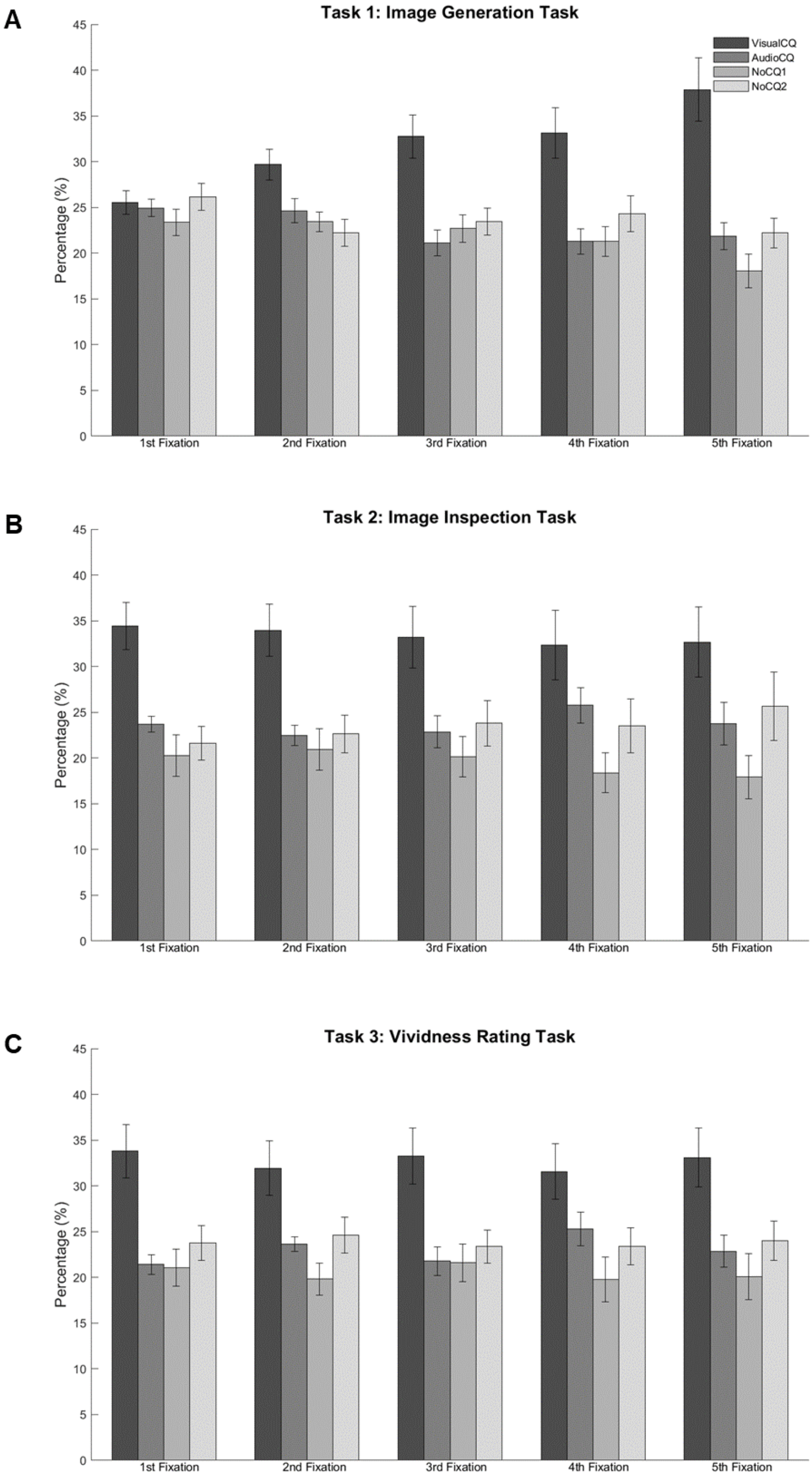

Figure 3. Graph showing the mean percentage of fixation in each quadrant from the first until fifth fixations during the Image Generation task (A), the Image Inspection task (B) and the Vividness Rating task (C) in the Imagery phase. Error bars represent ± 1 SEM.

It is interesting to see that in the Image Generation task the VisualCQ begins to influence eye behavior from the second fixation on and remains dominant across tasks. There is no evidence of an influence of the AudioCQ during the first fixations of mental image generation.

To rule out the possibility that participants were developing a retrieval strategy during the task (i.e. looking back to the areas where the images were located), we considered the order of trials. If participants were using a strategy, we would expect to see a change in the pattern of eye behavior when comparing first trials to later trials. During the first trials participants should still be naïve about the task, and for example not control eye behavior in an overt manner. It is conceivable that during the first trials eye movements might reflect the multiple representations of the ongoing process, thus including eye movements to the areas where the incongruent stimuli were presented. In Figure 4, we report dwell time in the different quadrants separated by order of trials and by task (Image Generation task, Image Inspection task, Vividness Rating task). Visual inspection of the graphs does not suggest a change in eye behavior during the unfolding of the task, i.e. when comparing first trials to later trials. We conclude that the AudioCQ did not trigger specific eye movements also at the beginning of the task.

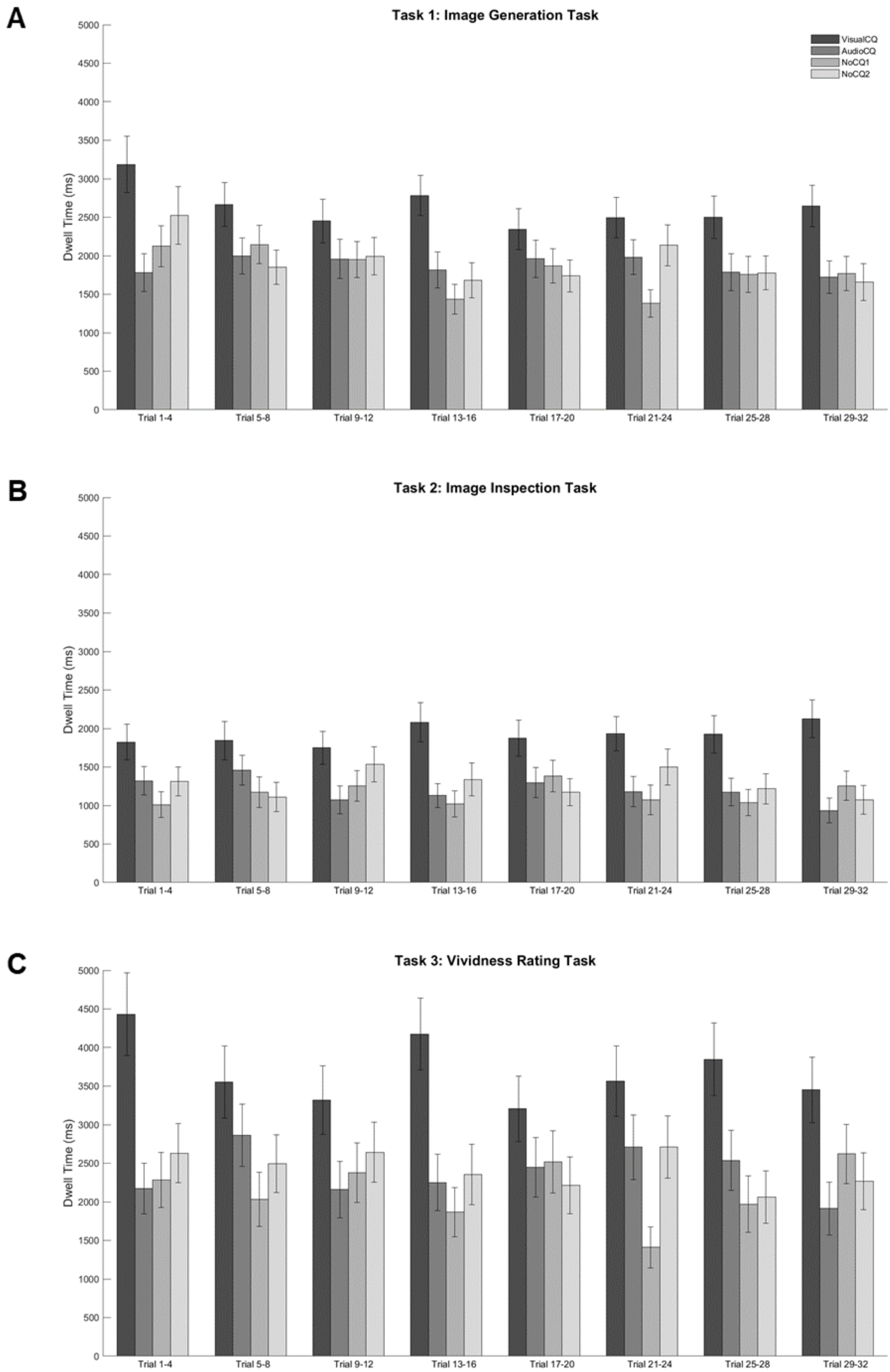

Figure 4. Graph showing the mean dwell time in each quadrant based on trial order during the Image Generation task (A), the Image Inspection task (B) and the Vividness Rating task (C) in the Imagery phase. Error bars represent ± 1 SEM.

Gender differences

Given the large overrepresentation of females ($n = 24$) over males ($n = 6$) in our sample, we run the mixed ANOVA presented in paragraph "Imagery phase: Percentage of dwell time in the three imagery tasks" by excluding the six men from the analyses. The results were highly similar. Future research should include a larger sample of males to investigate possible gender differences. Please see Appendix B for a full description of the analyses.

Discussion

We measured eye movements to understand how visual and auditory information is integrated when we create mental images. During encoding, participants spent significantly more time in the quadrant where the visual stimuli were located, as compared to the other quadrants, confirming participants' compliance. Task performance (accuracy) during imagery showed that participants did in fact use mental imagery. When participants were instructed to generate the mental image corresponding to a certain sound, participants looked longer in those areas in which that particular image had been previously seen. In contrast, they did not look longer in those areas in which they had seen another image associated to the same sound that cued the mental image. Whether the incongruencies were intra-categorical (e.g. an animal image paired with another animal sound) or extra-categorical (e.g. an animal image paired with an object sound) had no impact on eye behavior. Despite the audio-visual incongruencies during encoding, the re-enactment of gaze patterns during imagery relied exclusively on previous visuo-spatial associations of the to-be-imagined object.

Prioritization of visuo-spatial associations started from the beginning of the image generation process and was present in the first trials already. The original position associated with the sound was fully overruled by the location of the visual object. This finding suggests that the integrated memory trace is non-rigid and different aspects (i.e. visual, auditory, spatial, semantic) can be re-associated depending on the task at hand. Contrary to our expectations, we found no evidence for an automatic activation of eye movements toward those areas that were visited when participants heard the sound during encoding. The eye behavior suggests that incongruent visuo-spatial information that was possibly encoded along the auditory cues was not retrieved. Further, the results suggest that auditory cues were re-associated with the congruent visuo-spatial information presented in a close temporal context. Spatial correspondence between senses and coherent and stable representations are necessary for appropriate behavior (Berger & Ehrsson, 2018), and therefore it makes sense that visuo-spatial and auditory aspects are re-associated in a congruent manner in our study.

Participants looked significantly longer in those areas in which the image had been previously seen. These results suggest that during mental imagery participants were re-visiting and re-enacting the gaze pattern that they had shown when initially *perceiving* the image, looking longer where they had been previously looking. These results are in line with existing literature (e.g. Johansson, Holsanova, Dewhurst, & Holmqvist, 2012; Johansson & Johansson, 2014; Martarelli, Chiquet, Laeng, & Mast, 2017; Martarelli & Mast, 2011, 2013). The prioritization of visuo-spatial associations was true for all tasks (Image Generation task, Image Inspection task, and Vividness Rating task). It is interesting that participants continue to look back to the areas visited during encoding when answering questions about specific details of the image as well as when evaluating the vividness of the mental image. These results suggest that eye movements are equally relevant when generating a mental image, inspecting a mental image, and evaluating a mental image. It seems that eye movements occur not only as a reactivation of the oculomotor process in the first stages of the imagery process, but also then when the mental image has been created already (Laeng et al., 2014). Detailed mental images need continuous updating, and the reenactment process might support image generation as well as image maintenance. To our knowledge, this is the first study to show the looking back to nothing effect when judging the vividness of the mental image. Prior research focused on objective measures of mental imagery, such as true/false answers about specific details of the previously seen item (e.g. color).

Some limitations should be considered when interpreting the findings. A larger sample with females and males equally distributed should be tested. To allow a direct comparison in terms of visual and audio congruence, future research should also include congruent stimulus pairs as usually employed in the blank screen paradigm. Indeed, specific aspects of our task might have promoted visual input over auditory cues. For example, we used non-linguistic cues (e.g. Meeow) as stimulus material, whereas the blank screen paradigm is mostly used in combination with linguistic cues (e.g. cat). Moreover, the dominant visuo-spatial association that we found does not mean that participants do not have any memory of the incongruent audio-visual association presented at encoding. Fixations that are non-informative might simply not be reinstated. Future research should vary task instructions, for example by asking participants to remember the associated incongruent visual information or by using more episode-oriented tasks. This would help to better understand how visual and auditory information is integrated during the imagery process and elucidate the voluntary/involuntary components of eye behavior. Nevertheless, the looking back to nothing behavior appears to be a stable phenomenon, and it appears unaffected by the co-presence of linguistic information that was previously associated with other visuo-spatial information. Incongruent visuo-spatial information during encoding did not interfere with the mental imagery process. It seems that visual, auditory and spatial information is encoded in multiple traces that can be flexibly recombined during the imagery process depending on the demands of the ongoing task.

Compliance with ethical standards

Funding: This work was supported by a scholarship from the Ministry of Education Malaysia to HU (KPT(BS)850610065436).

Conflict of interest: The authors declare no competing interests.

Ethical approval: The study was in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

References

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*(4), 723-742.

Berger, C. C., & Ehrsson, H. H. (2018). Mental imagery induces cross-modal sensory plasticity and changes future auditory perception. *Psychological Science, 29*(6), 926-935.

Borst, G., & Kosslyn, S. M. (2008). Visual mental imagery and visual perception: Structural equivalence revealed by scanning processes. *Memory & Cognition, 36*(4), 849-862.

Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Jounal of Cognitive Neuroscience, 9*(1), 27-38.

Ferreira, F., Apel, J., & Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends in Cognitive Sciences, 12*(11), 405-410.

Hebb, D. O. (1968). Concerning imagery. *Psychological Review, 75*(6), 466.

Johansson, R., & Johansson, M. (2014). Look here, eye movements play a functional role in memory retrieval. *Psychological Science, 25*(1), 236-242.

Johansson, R., Holsanova, J., Dewhurst, R., & Holmqvist, K. (2012). Eye movements during scene recollection have a functional role, but they are not reinstatements of those produced during encoding. *Journal of Experimental Psychology: Human Perception and Performance, 38*(5), 1289.

Kumcu, A., & Thompson, R. L. (2018). Less imageable words lead to more looks to blank locations during memory retrieval. *Psychological Research*, 1-18.

Kumcu, A., & Thompson, R. L. (2016). *Spatial Interference and Individual Differences in Looking at Nothing for Verbal Memory.* Paper presented at the CogSci.

Lacey, S., & Lawson, R. (2013). *Multisensory imagery*: Springer Science & Business Media.

Laeng, B., Bloem, I. M., D'Ascenzo, S., & Tommasi, L. (2014). Scrutinizing visual images: The role of gaze in mental imagery and memory. *Cognition, 131*(2), 263-283.

Laeng, B., & Teodorescu, D.-S. (2002). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cognitive Science, 26*(2), 207-231.

Martarelli, C. S., Chiquet, S., Laeng, B., & Mast, F. W. (2017). Using space to represent categories: Insights from gaze position. *Psychological Research, 81*(4), 721-729.

Martarelli, C. S., & Mast, F. W. (2013). Eye movements during long-term pictorial recall. *Psychological Research, 77*(3), 303-309.

Martarelli, C. S., & Mast, F. W. (2011). Preschool children's eye-movements during pictorial recall. *The British Journal of Developmental Psychology, 29*(3), 425-436.

Richardson, D. C., Altmann, G. T., Spivey, M. J., & Hoover, M. A. (2009). Much ado about eye movements to nothing: A response to Ferreira et al.: Taking a new look at looking at nothing. *Trends in Cognitive Sciences, 13*(6), 235-236.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition, 76*(3), 269-295.

Ryan, J. D., & Shen, K. (2020). The eyes are a window into memory. *Current Opinion in Behavioral Sciences, 32*, 1-6.

Scholz, A., Klichowicz, A., & Krems, J. F. (2018). Covert shifts of attention can account for the functional role of "eye movements to nothing". *Memory & Cognition, 46*(2), 230-243.

Scholz, A., Mehlhorn, K., & Krems, J. F. (2016). Listen up, eye movements play a role in verbal memory retrieval. *Psychological Research, 80*(1), 149-158.

Viggiano, M. P., Giovannelli, F., Giganti, F., Rossi, A., Metitieri, T., Rebai, M., . . . Cincotta, M. (2017). Age-related differences in audiovisual interactions of semantically different stimuli. *Developmental Psychology, 53*(1), 138.

Appendix A: List of quadrant allocation for image-sound pairs used during the tasks.

In the intra-categorical group (Group A) as well as in the extra-categorical group (Group B) 32 stimuli were presented. Please find in the table below a list of stimuli used during the experiment as well as the position of the images on the screen. AOI 1 corresponds to the upper-left quadrant, AOI 2 corresponds to the upper-right quadrant, AOI 3 corresponds to the lower-right quadrant and AOI 4 corresponds to the lower-left quadrant.

Group A (Intra-Categorical pairs)			**Group B (Extra-Categorical pairs)**		
Quadrant	**Image**	**Sound**	**Quadrant**	**Image**	**Sound**
AOI 1	Dog	Hen	AOI 1	Cat	Car
AOI 4	Hen	Dog	AOI 4	Car	Cat
AOI 4	Tiger	Horse	AOI 4	Lion	Hand bell
AOI 1	Horse	Tiger	AOI 1	Hand bell	Lion
AOI 2	Pig	Frog	AOI 2	Monkey	Guitar
AOI 3	Frog	Pig	AOI 3	Guitar	Monkey
AOI 3	Elephant	Mosquito	AOI 3	Donkey	Police siren
AOI 2	Mosquito	Elephant	AOI 2	Police car	Donkey
AOI 1	Camel	Dolphin	AOI 1	Duck	Scissors
AOI 3	Dolphin	Camel	AOI 3	Scissors	Duck
AOI 3	Goose	Crocodile	AOI 3	Sheep	Vacuum
AOI 1	Crocodile	Goose	AOI 1	Vacuum	Sheep
AOI 2	Canary bird	Goat	AOI 2	Cow	Washer
AOI 4	Goat	Canary bird	AOI 4	Washer	Cow
AOI 4	Turkey	Bee	AOI 4	Wolf	Bagpipes
AOI 2	Bee	Turkey	AOI 2	Bagpipes	Wolf
AOI 1	Ship	Fire truck siren	AOI 1	Flies	Train
AOI 4	Fire Truck	Ship	AOI 4	Train	Flies
AOI 4	Piano	Trumpet	AOI 4	Snake	Airplane
AOI 1	Trumpet	Piano	AOI 1	Airplane	Snake
AOI 2	Helicopter	Motorcycle	AOI 2	Owl	Bicycle
AOI 3	Motorcycle	Helicopter	AOI 3	Bicycle	Owl
AOI 3	Violin	Bongos	AOI 3	Rooster	Tambourine
AOI 2	Bongos	Violin	AOI 2	Tambourine	Rooster
AOI 1	Whistle	Hair dryer	AOI 1	Eagle	Telephone
AOI 3	Hair Dryer	Whistle	AOI 3	Telephone	Eagle
AOI 3	Lawn Mower	Hammer	AOI 3	Mouse	Flute
AOI 1	Hammer	Lawn mower	AOI 1	Flute	Mouse
AOI 2	Alarm clock	Blender	AOI 2	Bear	Drum
AOI 4	Blender	Alarm clock	AOI 4	Drum	Bear
AOI 4	Printer	Microwave	AOI 4	Crickets	Electric drill
AOI 2	Microwave	Printer	AOI 2	Electric drill	Crickets

Appendix B: Gender differences

Given the large overrepresentation of females (n = 24) over males (n = 6) in our sample, we run the mixed ANOVA presented in paragraph "Imagery phase: Percentage of dwell time in the three imagery tasks" by excluding the six men from the analyses. The analysis again revealed a significant main effect of Cue Quadrant ($F(1.82, 40) = 8.00$, $p = .002$, $\eta^2_p = .27$). Dwell time was significantly higher in VisualCQ (35.6%, SD 16.8) than in the other Cue Quadrants (AudioCQ: 22%, *SD* 5.56, $p = .002$; NoCQ1: 19.3%, *SD* 9.22, $p < .001$; NoCQ2: 23.1%, *SD* 12.2, $p = .005$). However, in this analysis the interaction of Task with Cue Quadrant ($F(3.65, 80.33) = 2.75$, $p = .038$, $\eta^2_p = .11$) turned out to be significant. All other effects were non-significant. Please find in Table B1 a summary of the results of the ANOVAS for percentage of dwell time in the Imagery phase.

Table B1

Summary of the three-way ANOVAs for percentage of dwell time (Imagery phase) in the whole sample (n = 30) and in the female sample (n = 24).

Effects	***n* = 30**				***n* = 24**			
	df	***F***	***p***	η^2_p	***df***	***F***	***p***	η^2_p
Group (G)	1, 28	0.005	0.944	< .001	1, 22	0.001	> .05	< .001
Task (T)	1, 28	0.002	0.964	< .001	1, 22	0.002	> .05	< .001
Cue Quadrant (CQ)	**1.98, 55.37**	**7.410**	**0.001**	**0.209**	**1.82, 40**	**7.997**	**0.002**	**0.27**
G × T	1, 28	0.004	0.953	< .001	1, 22	0.003	> .05	< .001
G × CQ	1.98, 55.37	0.272	0.760	0.010	1.82, 40	0.652	0.512	0.029
T × CQ	3.83, 107.20	2.037	0.097	0.068	**3.65, 80.33**	**2.746**	**0.038**	**0.111**
T × CQ × G	3.83, 107.20	1.092	0.363	0.038	3.65, 80.33	1.415	0.240	0.06

Significant ANOVA terms are in bold. The degrees of freedom (*df*) were Huynh-Feldt corrected.

Given the significant interaction of Task with Cue Quadrant, we conducted three additional ANOVAs for each task on the percentage of dwell time spent in the quadrants. In the Image Generation task (Task 1), we found a significant effect of Cue Quadrant (*F*(2.32, 53.29) = 8.06, $p < .001$, $\eta^2_p = .26$). Post-hoc tests with Tukey correction showed that dwell time was significantly higher in VisualCQ (33%, *SD* 10.8) than in the other Cue Quadrants (AudioCQ: 22.8%, *SD* 4.84, $p = .002$; NoCQ1: 21.4%, *SD* 6.86, $p < .001$; NoCQ2: 22.8%, *SD* 8.53, $p = .002$). Similarly, in the Image Inspection task (Task 2), we also found a significant effect of Cue Quadrant (*F*(1.90, 43.81) = 7.08, $p = .002$, $\eta^2_p = .24$) and post-hoc tests with Tukey correction showed that dwell time was significantly higher in VisualCQ (37.6%, *SD* 19.4) than in the other Cue Quadrants (AudioCQ: 21.7%, *SD* 6.24, $p = .004$; NoCQ1: 19%, *SD* 11.4, $p < .001$; NoCQ2: 21.7%, *SD* 13.8, $p = .004$). Finally, in the Vividness Rating task (Task 3), we also found a significant effect of Cue Quadrant (*F*(1.62, 37.26) = 6.89, $p = .005$, $\eta^2_p = .23$). Post-hoc tests with Tukey correction showed that dwell time was significantly higher in VisualCQ (36.1%, *SD* 19.3) than in the AudioCQ (21.4%, SD 5.64, $p = .006$) and NoCQ1 (17.5%, *SD* 8.87 $p < .001$). The difference between VisualCQ and NoCQ2 (25%, *SD* 13.8, $p = .056$) just missed significance. The pattern of results turned out to be very similar, when excluding males from the sample. Future research should include a larger sample of males to investigate possible gender differences.