



HOKKAIDO UNIVERSITY

Title	Commonalities of visual and auditory working memory in a spatial-updating task
Author(s)	Maezawa, Tomoki; Kawahara, Jun I.
Citation	Memory & cognition, 49(6), 1172-1187 https://doi.org/10.3758/s13421-021-01151-8
Issue Date	2021-02-22
Doc URL	https://hdl.handle.net/2115/84202
Rights	This is a post-peer-review, pre-copyedit version of an article published in Memory & cognition. The final authenticated version is available online at: http://dx.doi.org/10.3758/s13421-021-01151-8 .
Rights(URL)	https://www.springer.com/gp/open-access/publication-policies/self-archiving-policy
Type	journal article
File Information	Commonalities_of_visual_and_auditory_working_memory.pdf



Commonalities of visual and auditory working memory in a spatial-updating task

Tomoki Maezawa and Jun I. Kawahara

Department of Psychology,
Hokkaido University, Sapporo, Japan

Running head: Common visual and auditory spatial working memory

Corresponding author

Tomoki Maezawa

Hokkaido University,

N10W7, Kita

Sapporo, 060-0810, Japan

Tel: +81 11-706-4198

E-mail: porpoise@let.hokudai.ac.jp

Keywords: Working memory, Audition, Spatial updating, Imagery processing, Location memory

Abstract

Although visual and auditory inputs are initially processed in separate perception systems, studies have built on the idea that to maintain spatial information these modalities share a component of working memory. The present study used working memory navigation tasks to examine functional similarities and dissimilarities in the performance of updating tasks. Participants mentally updated the spatial location of a target in a virtual array in response to sequential pictorial and sonant directional cues before identifying the target's final location. We predicted that if working memory representations are modality-specific, mixed-modality cues would demonstrate a cost of modality switching relative to unimodal cues. The results indicate that updating performance using visual unimodal cues positively correlated with that using auditory unimodal cues. Task performance using unimodal cues was comparable to that using mixed modality cues. The results of a subsequent experiment involving updating of target traces were consistent with those of the preceding experiments and support the view of modality-nonspecific memory.

Introduction

The spatial locations of objects are perceived through separate visual and auditory systems. The inputs from different sensory modalities are processed in separate primary stages in retinotopic and tonotopic areas for visual and auditory information, respectively. The resulting representations differ in terms of spatial resolution (Alais & Burr, 2004; Cohen et al., 2009) and the reference frames of spatial coordinates (Colby & Goldberg, 1999; Goossens & Van Opstal, 1999). Specifically, the spatial components of visual inputs are encoded in greater spatial detail than are auditory inputs and are represented in terms of eye-centered coordinates, whereas auditory inputs are represented by head-centered coordinates. The differences between visual and auditory representations continue beyond perception. The traces of auditory sensory memory (echoic) persist longer than do those of visual sensory (iconic) memory (Craik, 1969; Javitt et al., 1997; Sperling, 1960). For example, recognition memory for non-verbal visual objects is superior to auditory recognition memory, even though visual stimuli do not consist of more information than auditory stimuli (Cohen et al., 2009, 2011).

Because of the aforementioned differences in visual and auditory representations, researchers have examined whether spatial information is maintained in separate or shared modality-specific systems. A multisensory perspective of working memory is

relevant to examining these questions (for a review, see Quak et al., 2015).

Working memory has been referred to as a temporary storage system for the mind to manipulate information during concurrent tasks (Baddeley, 2012; Cowan, 2016; Ricker et al., 2010). The process of working memory involves encoding, maintaining, and updating the contents in storage. One current theory of working memory conceptualizes a multiple-component model that proposes domain-specific slave systems of the central executive (Cowan, 1999; Hitch et al., 2020). According to this model, two subsidiary components, the visuo-spatial sketchpad and the episodic buffer, are associated with maintaining and updating mental images (Baddeley, 2002; Baddeley et al., 2011) via attention control (Hitch et al., 2020; Logie, 1995, 2011). For example, the visuo-spatial sketchpad allows rehearsal of a short sequence of movements and paths between locations, which can be stored and manipulated by focusing and sustaining attention (Logie, 2011). Evidence supporting these theories has been obtained mainly from studies using visual materials (e.g., Luck & Vogel, 1997, 2013; Mance & Vogel, 2013). Thus, greater emphasis has been placed on the role of the visual domain in studies of spatial working memory. Understanding of the maintenance of audio-spatial information is incomplete.

Studies of audio-spatial working memory (e.g., Alain et al., 2008, 2009; Delogu et

al., 2012; Kaiser, 2015; Lewald & Ehrenstein, 2001; Martinkauppi et al., 2000; Vuontela et al., 2003) have been built on two models of the relationship between audio- and visuo-spatial buffering components. In the first model, working memory consists of separate auditory-specific buffering components that represent objects' spatial codes temporarily in relation to the visuo-spatial component. In the second model, the buffering components of working memory are shared by vision and audition. These two models are underpinned by evidence of modality-specific and shared spatial attention (Kong et al., 2014; Lehnert & Zimmer, 2008a; Michalka et al., 2015). Because spatial attention is necessary for optimally functional spatial working memory systems (see Awh et al., 1998; Baddeley, 2002; Smyth & Scholey, 1994), it is important to determine whether the visual and auditory systems use separate or shared attentional resources (Lehnert & Zimmer, 2008b). For example, the idea of attentional resources posits separate or shared components in working memory that store types of spatial information against the sensory modalities. For decades, evidence has accumulated that favors both the separate (Bushara et al., 1999; Kong et al., 2014; Michalka et al., 2015; Sinnott et al., 2007) and shared (Driver & Spence, 2004; Krumbholz et al., 2009; Lehnert & Zimmer, 2008a; Martinkauppi et al., 2000) views of attention. Therefore, the concept of audio-spatial working memory can be associated with either theoretical

position.

Despite the theoretical possibility of separate working memory (e.g., Tremblay et al., 2006), a limited number of studies on audio-spatial working memory imply a contribution of common components in the maintenance of spatial contents (Martinkauppi et al., 2000; Lehnert & Zimmer, 2006, 2008a, 2008b; Loomis et al., 2012). These studies used an experimental paradigm similar to conventional visuo-spatial tasks (e.g., Luck & Vogel, 1997), in which participants stored a sequence of location-bound stimuli, and subsequently recognized the location of the test stimulus. Importantly, research on audio-spatial working memory has aimed to evaluate and compare recognition performance both in a given single modality and in a mixture of visual and auditory modalities. In a series of studies by Lehnert and Zimmer (2006, 2008a), sequential stimuli consisted of audiovisual items that were all visual, all auditory, or a mixture of the two. In the mixture condition, switches between the two modalities occurred during the trial. Lehnert and Zimmer (2006) found no reliable effect of switching modalities on working memory capacity, indicating that the spatial-related components of working memory may be shared by vision and audition.

Audio-spatial working memory studies (Martinkauppi et al., 2000; Lehnert & Zimmer, 2006, 2008a; Loomis et al., 2012) have explored the commonality of working

memory in storing information about the location of items. However, aside from these studies, there is a clear functional and neurophysiological distinction between the two working memory processes of maintaining (e.g., Vogel et al., 2005) and updating (e.g., Bledowski et al., 2004; Polich, 2007). The process of updating mental representation has been identified as being a higher cognitive function that compares and integrates new inputs with previous contents in working memory. This updating may cause performance deficits when the contents consist of different modality forms and interfere with one another (e.g., Kreutzfeldt et al., 2017; Loomis et al., 2012). The cost is concerned with the modality-specific influences on attention and working memory function. One possibility is that separation of spatial inputs leads to modality-specific labels remaining in representations, which interfere with each other (Loomis et al., 2012). Additionally, there are costs in task performance when participants shift their attention from items of one modality to items of a different modality (Kreutzfeldt et al., 2017; Ruthruff et al., 2001; Spence et al., 2001; Suchan et al., 2006). Task switching requires an effort to prepare for the upcoming task, which imposes a greater load on working memory (Kiesel et al., 2010; Monsell, 2003). Such detection of the cost of modality switching during updating is in sharp contrast to the findings of studies that focused on the maintenance function. Because modality-switching might incur higher

demands and costs to update than to maintain items, evidence from studies focused on the maintenance function of working memory (Lehnert & Zimmer, 2006, 2008a) may be asymmetric to that from studies focused on the updating function.

The present study examined functional similarities or dissimilarities in working memory in response to visuo- and audio-spatial cues during updating spatial representation. In the first experiment, we focused on the correlation between visual and auditory task performance. In subsequent experiments, we focused on the performance cost of modality-switching in terms of response accuracy while updating working memory. Each experiment consisted of a mental-pathway task (Cornoldi et al., 1991; Kerr, 1993) in which participants were asked to mentally maneuver a representation of a target in an imaginary space using visual and/or auditory directional cues, and to follow the target's pathway to its final destination. Visuo- and audio-spatial materials were used during the task. In this trial, participants maintained the current location of the target and updated it following a new directional input. The present study aimed to measure whether performance of the spatial updating task in the single visual or auditory modality was superior to that in the mixed modality. Across the four experiments, we predicted two main findings, in accordance with the modality-general updating process: there would be a positive correlation between visual and auditory task performance

(Experiment 1), and task performance in the context of mixed cues would not be impaired relative to that for single modality cues, i.e., no modality-switching cost (Experiments 2 and 3).

Experiment 1

Prior to testing the cost of modality-switching using mixed sensory inputs, Experiment 1 examined response accuracy in a visuo-spatial task and that in an audio-spatial task. These tasks were designed to compare accuracy directly and determine similarities or dissimilarities. During this experiment, participants maneuvered a spatial representation of a target to follow a mental pathway to its final destination in response to directional cues indicating movement right, left, up or down. This method included two types of directional cues in two separate modalities (vision and audition). The discriminability of the auditory stimuli was lower than that of the visual stimuli. This difference may have been due to a difference in spatial resolution between the visual and auditory stimuli, which affects the early stages of working memory processing such as encoding (for a review, see Lehnert & Zimmer, 2008b). If this were the case, performance on the single visual matrix task in Experiment 1 would be higher than that on the single auditory task, as reported in a previous study on the capacity limitations of auditory working memory (Lehnert & Zimmer, 2006). The superiority of the visual over

the auditory modality reflects the preference for visual stimuli during the early stages of working memory processes, rather than the post-perceptual updating process.

Regardless of whether there are discriminability differences between the two modalities, a strong positive correlation should be detected if the updating of different types of stimuli in spatial working memory relies on a non-modality-dependent system (also see Loomis et al., 2012). This prediction was based on the idea that high (low)-capacity visual spatial updating is associated with high (low)-capacity auditory spatial updating. The correlation might be reduced if the superiority of either visual or auditory performance of a spatial task varied across participants. In this situation, performance under the visual modality would not be associated with that under the auditory modality and vice versa. The asymmetry of the performance would reflect separate working memory systems.

Method

Participants. A priori power analysis was conducted using an online calculator in PANGEA (Westfall et al., 2014) based on a within factorial design that consisted of $f = 0.25$ of the effect size with a sample size of 30. The analysis achieved 0.8 power to detect the main effects of the modalities of directional cues. A group of 30 naive students from a participant pool at Hokkaido University (eight female

participants; mean age = 20.1 years; range: 18–23 years) was recruited and compensated by either monetary payment or course credit. All participants reported normal or corrected-to-normal visual acuity. None of the participants reported any hearing impairment. All participants provided written informed consent prior to the experiment in accordance with the requirements of the Human Research Ethics Committee of Hokkaido University.

Visual stimuli. The experiment was run in a well-lit room. Visual stimuli were presented on a black background on a 24-inch LCD monitor (XL2411T; BenQ Co., Ltd., Taiwan; 100 Hz refresh rate, $1,920 \times 1,080$ pixels) controlled by MATLAB (The MathWorks, Inc., USA) using the Psychophysics Toolbox (Kleiner et al., 2007). The viewing distance between the display and the participant's eyes was approximately 57 cm. Responses were collected using a computer mouse and keyboard. During the mental-pathway task, a white unfilled matrix of 11×11 cells ($2.2^\circ \times 2.2^\circ$) was displayed on the screen. The initial location of the target stimulus was indicated as a filled red square ($2.2^\circ \times 2.2^\circ$) in the matrix. The visual-modality directional cues were images of white arrows ($9^\circ \times 9^\circ$) indicating one of four directions (right, left, up, or down), in response to which participants were supposed to move the target.

Auditory stimuli. Auditory stimuli were presented through headphones (MDR-

XB550; Sony Co., Ltd., Japan) via a PC/AT-compatible Linux-operated computer (OptiPlex 990; Dell Inc., USA) that was equipped with a standard audio chipset (ALC269; Realtek Semiconductor Corp., Taiwan). All auditory stimuli were generated using MATLAB. Auditory directional cues were based on white Gaussian noise of 1,000 ms duration digitized at 44.1 kHz with 16-bit depth. To generate these cues, we applied sound specialization to simulate virtual sound sources on the surface of a sphere on which a triple-coordinate (x, y, z) system was applied. First, the sound file of the noise was converted into a first-order ambisonics format with SN3D normalization using a plugin (Oculus Spatializer; Oculus VR, USA) for a digital audio workspace (REAPER; Cockos, Inc., USA). Second, binaural decoding was conducted using a filter of head-related transfer functions (HRTFs) with MATLAB code. The filter was applied to a subset of the ARI database (Acoustics Research Institute, Austria; retrieved from <https://www.oeaw.ac.at/en/isf/das-institut/software/hrtf-database>). Each auditory spatial cue was assigned to one of four corresponding virtual sound sources (Figure 1A; right, left, up, or down). Specifically, the right and left sources corresponded to right and left locations on the horizontal plane 1.00 m away from the subject's head. The up and down positions corresponded to locations on the sagittal plane 0.60 m anterior and 1.04 m superior relative to the head position for the up position, and 2.60 m posterior and

1.50 m inferior relative to the head position for the down position. Regardless of the direction, the cues were generated at a constant sound level at the source position.

Procedure.

Participants completed the mental-pathway task under both the visual and auditory modality experimental conditions. The two modalities were manipulated blockwise. The visual block included only a working memory task, while the auditory block included a pretest of hearing the virtual cues followed by the main working memory task. The order of administration of modality blocks was counterbalanced across participants.

A hearing pretest was conducted before the auditory mental-pathway task detailed below. The purpose of this pretest was to familiarize participants with hearing the virtual sounds and to test their ability to discriminate the direction of these sources in terms of right, left, up, and down. The pretest was divided into practice and test sessions. In the practice session, the auditory directional cues were presented six times per direction, proceeding clockwise and starting from the right. In the following test session, the auditory cues were presented pseudo-randomly, and participants identified the direction of the cues using a computer keyboard by pressing the 'd', 'a', 'w' or 's' key. Feedback was provided when a participant failed to identify the direction correctly and a dot indicated the correct direction. Participants completed a total of 48 trials that

consisted of 12 trials per direction. The percentage of correct responses was calculated for each participant. In the hearing pretest, the mean accuracy of the 30 participants was 98.82% ($SD = 0.01$) across the four directions in response to the auditory stimuli (vs. 100 % accuracy for visual stimuli), beyond the level of chance (= 25%).

This study extended a conventional measure of working memory function (Cornoldi et al., 1991; Kerr, 1993), in that the task was able to assess both visuo-spatial and audio-spatial performance. Figure 2 shows the procedure of a single trial. Each trial started with presentation of a white central fixation cross ($0.2^\circ \times 0.2^\circ$) for 1,000 ms. The 11×11 matrix ($24.2^\circ \times 24.2^\circ$) with the target was then presented for 5,000 ms. The initial location of the target was pseudo-randomly selected within the 121-cell matrix. After a blank of 1,000 ms, the visual or auditory directional cues were sequentially presented on the screen or via headphones for 1,000 ms, respectively. Each cue was followed by a visual or auditory blank period of 1,000 ms. Then, the matrix was displayed again until the participant's response. This matrix display did not contain the initial location of the target and served as a test screen.

Participants were instructed to memorize the initial location of the target, and subsequently maneuver the target in response to the cued direction by reference to spatial imagery. We also instructed them not to use strategies such as verbal labeling

and/or finger pointing during the navigation task. Eventually, they indicated the final location of the target by clicking on the corresponding cell in the matrix. The sequence length of directional cues dictates the working memory load associated with navigation. The sequence length varied substantially among previous studies, from 2 to 12 items (Attneave & Curlee, 1983; Kerr, 1993; Tao et al., 2017), which may have been responsible for differences in performance among those studies. To examine the effect of working memory load, in Experiment 1 the sequence length was varied (four, six, or eight) in the matrix task. The number of directional cues in a sequence was pseudorandomized across trials. Participants completed two blocks of a total of 60 trials that consisted of two modality blocks with 30 trials in each block and yielded 10 trials for each condition of modality \times sequence length as within-subject factors. Before undertaking the 60 trials, participants were familiarized with the task through three practice trials at the beginning of the experimental trials.

Statistical analysis. The present study considered rejection of a hypothesis based on the idea of separation of spatial working memory. Accordingly, we investigated possible differences in working memory associated with different sensory modalities. In Experiment 1, we examined the functional similarity or dissimilarity of participants' performance using a single modality, either visual or auditory. However, the framework

of null-hypothesis statistical testing does not support data under the null hypothesis (Dienes & Mclatchie, 2018). To circumvent this limitation, we used Bayesian hypothesis testing (Rouder et al., 2017; Wetzels et al., 2012) with estimation of Bayes factors (BF_{10} ; Morey & Rouder, 2011). BF_{10} represents the likelihood of one hypothesis (e.g., alternative) being favored over another hypothesis (e.g., null); specifically, BF_{10} quantifies the degree to which the data support either hypothesis. According to a common classification (Jeffreys, 1961), a BF_{10} value of 1 indicates *no* evidence for either hypothesis, while values below 1 and above 1 support the null hypothesis and the alternative hypothesis, respectively; BF_{10} values below 0.33 (= 1/3) and above 3 provide a *moderate* degree of evidence in support of the null hypothesis and the alternative hypothesis, respectively; and BF_{10} values below 0.10 and above 10 provide *strong* evidence for the respective hypothesis. Note that $0.33 < BF_{10} < 1$ and $1 < BF_{10} < 3$ represent *inconclusive* evidence rather than *no* evidence. In this study, confirmation of our prediction favoring common working memory hinges on the BF_{10} . A value below 0.33 favors the null hypothesis while one above 3 favors the alternative hypothesis.

Accuracy in the mental-pathway task was represented by the percentage of times the final destination of the target was identified correctly. The pooled percentage of correct responses was calculated by averaging participants' scores over the experimental

factors of modality \times sequence length per participant. To evaluate the associations between the performances under the two modalities, correlations were analyzed in the manner of Bayesian analyses. BF_{10} values were computed using JASP software (JASP Team, 2019) with default priors as implemented in JASP (see Morey et al., 2018; Wagenmakers et al., 2018). We conducted a Bayesian repeated-measures ANOVA for the percentages of correct responses with modality (visual or auditory) and sequence length (four, six, or eight) as within-subject factors. The ANOVA results were analyzed using post-hoc tests by the Bayesian method.

Results

Figure 3 displays a plot of the mean percentages of correct responses for the visual and auditory modalities as a function of sequence length (four, six, or eight). The ANOVA revealed that an effect of modality was inconclusive in terms of whether the null hypothesis was supported, $BF_{10} = 0.47$ (i.e., > 0.33). The main effect of sequence length favored the alternative hypothesis, $BF_{10} = 2.96 \times 10^2$ (> 3). The two-way interaction term supported the null hypothesis, $BF_{10} = 0.21$ (< 0.33). Post-hoc tests showed that the mean percentage of correct responses with a sequence length of four was greater than that with a sequence length of eight, $BF_{10} = 7.09 \times 10^2$. The comparisons also yielded a $BF_{10} < 3$, which is inconclusive in terms of supporting the

alternative hypothesis of a difference between the sequence lengths of four and six, $BF_{10} = 1.86$, and six and eight, $BF_{10} = 2.54$.

The percentage of correct responses averaged across sequence lengths was subjected to a Bayesian correlation analysis to assess the similarity between visual and auditory task performance. Figure 4 shows that visual task performance across sequence lengths was strongly correlated with auditory task performance, $r(30) = .718$, $BF_{10} = 3.04 \times 10^3$, supporting the alternative hypothesis.

Discussion

The primary finding of this study was that there was a strong positive correlation between participants' performance in the visuo-spatial and audio-spatial tasks. This result implies that after encoding, visual and auditory stimuli are involved in a similar process of updating spatial representation. Thus, the present experiment suggests that high performance on visual (auditory) spatial updating processes would predict high performance on auditory (visual) spatial updating processes.

Analysis of the main effect of modality is critical for determining its impact on spatial task performance. The ANOVA yielded a BF_{10} that did not conclusively support the alternative or null hypothesis. This result was unexpected, given the poor discriminability of auditory cues (mean accuracy was below 100%) than visual cues.

However, high navigation accuracy (beyond the level of chance) might reduce the likelihood of detecting differences between the modalities. In fact, the present study replicated the finding that task performance tends to be better in mental-pathway tasks using a two-dimensional matrix (Cornoldi et al., 1991). Thus, further experiments are needed to confirm the effect of modality.

Experiment 1 focused on the effect of sequence length. As indicated by the ANOVA results, navigation accuracy decreased as sequence length increased, reflecting a shortage of working memory resources. The performance under the four-item sequence length condition tended to be better than that for the other lengths, consistent with the view that working memory can maintain representations of up to four items simultaneously. Beyond that threshold, the impact of increasing the sequence length on performance would be negligible due to a floor effect related to a shortage of working memory resources, as shown in the present study for the sequence lengths of six and eight. In accordance with Lehnert and Zimmer (2006), the effect of working memory load was independent of modality, such that the results for the interaction of modality \times sequence length favored the null hypothesis.

Therefore, the aim of Experiment 1 was to determine whether performance on a single-modality task (e.g., visual) was correlated with that for a task based on another

modality (e.g., auditory). However, as we were mainly interested in the modality-switching cost of task performance, a comparison of the participants' accuracy in response to the single and mixed modalities was needed for our second experiment. The existence of a switching cost is a vital indicator of similarities or dissimilarities between the visual and auditory working-memory systems, because such a cost would reflect a modality-specific effect on maintaining and updating spatial representations (e.g., Loomis et al., 2012). By contrast, if the two modalities share working memory, then information processing can be faster and more accurate than when the components are separated across sensory modalities. Thus, in Experiment 2 we used the same procedure as in Experiment 1, but included a sequence of stimuli with mixed modalities in the mental-pathway.

Experiments 2A and 2B

Experiments 2A and 2B included a trial block that comprised a sequential mixture of visual and auditory materials in a spatial working memory task. In the mixed condition, the modality of the sequence of visual and auditory directional cues was switched during the trial. Participants did not expect successive modalities in the sequence and the modalities were switched pseudo-randomly. Experiment 2A followed a similar procedure to that used in Experiment 1. Experiment 2B was almost identical to

Experiment 2A, except that a diagonal coordinate system was used such that the stimuli were orientated at 45° to the craniocaudal axis. Experiment 2B was expected to produce an oblique effect that would decrease participants' ability to mentally maneuver the target's location. This manipulation of rotating the display resulted in poorer navigation performance in the present experiment relative to the previous one.

In Experiments 2A and 2B, each directional cue was separated by a blank interval of 1,000 ms. This duration was less than that used in previous studies (e.g., Loomis et al., 2012; Suchan et al., 2006). When successive inputs were separated by 1.0 s, Loomis et al. (2012) found performance deficits in terms of response latencies when different modalities immediately followed one another at the same location. Experiments 2A and 2B in the present study were designed to examine whether there was a cost in switching modalities during working memory navigation. We predicted that there would be no performance deficits under mixed modalities relative to a single modality.

Method

Participants, apparatus and stimuli. The sample included the same number of participants as Experiment 1 and all participants passed the hearing pretest detailed below. One new group of 30 students (14 female participants; mean age = 19.2 years; range: 18–22 years) participated in Experiment 2A and another new group of 30

students (nine females; mean age = 20.0 years; range = 18–24 years) participated in Experiment 2B.

The apparatus and stimuli in Experiment 2A were identical to those used in Experiment 1. The visual and auditory stimuli in Experiment 2B were orientated at 45° to the craniocaudal axis. The differences between Experiments 2A and 2B were whether the visual and auditory cues indicated cardinal (i.e., right, left, up, or down for Experiment 2A) or diagonal (i.e., up-right, up-left, down-right, or down-left for 2B) directions. The visual stimuli in Experiment 2B were configured as an 11 × 11 matrix oriented at 45° and similarly oriented arrows. The auditory directional cues were a corresponding set of simulations of the four diagonal virtual sound sources (Figure 1B). Specifically, the up-right source position corresponded to the location on the coronal plane of 0.60 right and 1.04 superior relative to the head position and the up-left source position corresponded to 0.60 left and 1.04 superior relative to the head position. The down-right and down-left source positions corresponded to 2.60 right and 1.50 inferior and 2.60 left and 1.50 inferior relative to the head position, respectively.

Procedure and statistical analysis. A hearing pretest was administered before the auditory block of the mental-pathway tasks in Experiments 2A and 2B. The procedure was identical to that used in Experiment 1, except for the keys used to enter

responses in Experiment 2B. Specifically, the participants pressed keys ‘k’, ‘s’, ‘m,’ or ‘x’ for the up-right, up-left, down-right, or down-left directions, respectively. The mean accuracies of the 30 participants across the four directions in the hearing pretest were 98.82% ($SD = 0.01$) in Experiment 2A and 98.26% ($SD = 0.02$) in Experiment 2B. An additional five subjects participated in the Experiment 2B pretest (one female participant; mean age = 20.4 years; range: 18–23 years); their mean accuracy across the four directions was 87.50% ($SD = 0.08$). Because their accuracy for each direction was below 90%, they were excluded from participation in the main mental-pathway task and thus from the analyses.

Figure 5 shows a schematic diagram of a single trial in the mental-pathway tasks of Experiments 2A and 2B. The experiments constituted three blocks, with the visual-single, auditory-single, and mixed-modality blocks administered in that order. The sequence of the blocks was not counterbalanced across participants, because whether there was an order effect on the percentage of correct responses in Experiment 1 was inconclusive ($BF_{10} = 2.65$). The single-modality conditions were almost identical to those of Experiment 1. During the mixed-modality block, the modality of the directional cues could be switched randomly within a single trial (e.g., from auditory to visual or vice versa, etc.). We omitted the lowest memory load condition (four items) from

Experiments 2A and 2B, to avoid a ceiling effect with respect to performance and to enable the data to be collected within a reasonable testing period (approximately 60 minutes); this allowed 12 trials to be completed per condition. The sequence length (six or eight) was determined within each block. Participants completed a total of 72 experimental trials that consisted of three blocks with 24 trials in each block, yielding 12 trials for each condition of modality \times sequence length of the within-subject factors. All participants performed two practice trials at the beginning of the experimental trials to become familiar with the task. The percentage of correct responses per participant was calculated for each experimental factor. We conducted an ANOVA and post-hoc tests for a correct response with modality (visual, auditory, or mixed) \times sequence length (six or eight) as factors.

Results and Discussion

Experiment 2A. Figure 6A displays the mean percentages of correct responses as a function of sequence length (six or eight), separately, for the three modalities (visual, auditory, or mixed) in Experiment 2A. The ANOVA revealed a main effect of modality, $BF_{10} = 18.20$, and sequence length, $BF_{10} = 17.29$. The interaction term supported the null hypothesis, $BF_{10} = 0.17$, similar to Experiment 1. Post-hoc tests revealed that the mean percentage of correct responses under the visual modality was

greater than those under the auditory modality, $BF_{10} = 4.48$, and the mixed modality, $BF_{10} = 20.29$. Importantly, no difference was found between the auditory and mixed modality conditions, $BF_{10} = 0.17$.

Experiment 2B. Figure 6B displays the mean percentages of correct responses as a function of sequence length for the three modalities in Experiment 2B. The ANOVA revealed a main effect of modality, $BF_{10} = 3.60 \times 10^3$. The result regarding the effect of sequence length was inconclusive in terms of providing support for the alternative hypothesis, $BF_{10} = 2.84$. The interaction term was in favor of the null hypothesis, $BF_{10} = 0.19$. These results can be attributed to limited working memory capacity, irrespective of task modality. Post-hoc tests revealed that the mean percentage of correct responses under the visual modality was greater than those under the auditory modality, $BF_{10} = 66.43$, and mixed modality, $BF_{10} = 7.39 \times 10^2$. Post-hoc tests revealed a favor of no difference between the auditory and mixed modality, $BF_{10} = 0.16$.

To examine the oblique effect, we further conducted an ANOVA on the mean percentages of correct responses as a mixed between-within subject factor of experiment (Experiment 2A or 2B), modality (visual, auditory, or mixed), and sequence length (six or eight). The results revealed a main effect of experiment, $BF_{10} = 86.41$, modality, $BF_{10} = 1.18 \times 10^6$, and sequence length, $BF_{10} = 2.17 \times 10^2$. Additionally, the

interaction terms favored the null hypothesis, experiment \times modality, $BF_{10} = 0.13$, experiment \times sequence length, $BF_{10} = 0.16$, modality \times sequence length, $BF_{10} = 0.07$, and three-way interaction, $BF_{10} = 0.32$. Post-hoc tests revealed that the mean percentage of correct responses under the visual modality were greater than those under the mixed modality, $BF_{10} = 5.47 \times 10^4$, and the auditory modality, $BF_{10} = 1.83 \times 10^3$. No difference between the auditory and mixed-modality condition was observed, $BF_{10} = 0.13$.

Single visual modality performance was clearly better in Experiments 2A and 2B than single auditory modality performance, in contrast to the results of Experiment 1 (wherein there was no clear performance difference between modalities). These results are consistent with the findings of Lehnert and Zimmer (2006, 2008a), who suggested that the quality of visual and auditory inputs inherently differs due to differences in their spatial resolution (Alais & Burr, 2004; Cohen et al., 2009). It should be noted that the possible perceptual differences should also have impaired the mixed-modality performance to the same degree as auditory navigation, compared to the visual modality, given that the same (poor) auditory directional cues were used during the trials. Thus, superior performance in the auditory domain over the mixed one would reflect a modality-switching cost in Experiments 2A and 2B.

In fact, the results of Experiments 2A and 2B consistently indicated no difference in navigation accuracy between the auditory and mixed modalities. This finding was critical in the present study assessing a possible modality-switching cost. If there was a cost, the mixed-modality performance should have declined more than the performance in response to the single auditory modality. This was not the case; the results indicated no deficit as a result of modality switching.

The results of Experiments 1 and 2 favored functional similarities between visuo- and audio-spatial working memory, based on the correlation coefficients and modality-switching cost. There are a few explanations for the modality-switching cost discussed above, which support the modality-specificity of spatial working memory. The present results support a recoding strategy (Lehnert & Zimmer, 2006; 2008a), in which auditory-specific information is converted into visual format and then processed in visual working memory, together with visual information. This recoding process may be similar to the transformation process in verbal working memory whereby visual inputs are recoded into auditory format (Smith & Jonides, 1997; Suchan et al., 2006).

According to this perspective, conditions with a single auditory cue inherently involve a modality switch between the visually encoded information and the perceived auditory cue. Thus, the better performance in the visual condition than the auditory condition

was intuitive. Moreover, the difference in performance between the visual and auditory modalities in the current study support a modality-switching cost.

The results of Experiments 1 and 2 did not rule out a strategy based on verbal labeling during navigation tasks. Participants can use verbal statements instead of spatial imagery to code the position of the target as a number, similar to the position in the matrix. For instance, the number “23” refers to a location in the second row and third column of the matrix. Participant maneuvered the target by changing its number based on the cue, irrespective of the modality; thus a non-modality switch cost can be expected. In fact, Cornoldi et al. (2009) demonstrated that participants could use verbal strategies other than spatial imagery in the matrix task.

The present results were limited by differences in the discriminability of cues between modalities. Because mixed-cue trials contained a mixture of visual and auditory cues, performance was expected to fall between that of visual and auditory conditions. However, the results indicated that the performance under the mixed condition was comparable to that under the auditory condition. The lack of a difference between the auditory and mixed modalities in the present study supports a modality-switching cost.

Experiment 3 was designed to rule out alternative explanations for the results of

Experiment 2. This experiment manipulated the discrimination threshold of the visual cues by implementing parameter estimation (sequential testing, PEST; see Hall, 1974; Taylor & Creelman, 1967). Specifically, using the PEST staircase, we determined the thresholds whereby the modalities are comparable in terms of the discriminability of directional cues. Experiment 3 included three conditions: single visual, single auditory, and a mixture of the two. Accuracy under the visual and auditory modalities together indicate whether the visual recoding strategy described above was used. We examined whether the accuracy was lower in the single auditory condition than that in the single visual condition. To rule out an explanation based on verbal labeling, Experiment 3 involved articulatory suppression, to test whether performance under the mixed condition is comparable to that under single visual and auditory conditions. We expected that there would be no performance difference between the mixed and single modalities, based on the principle that working memory is not modality-specific.

Experiment 3

The primary aim of Experiment 3 was to exclude alternative explanations for the results of Experiment 2. One possibility was that the superior performance in the visual versus auditory domain seen in Experiment 2 might reflect the cost of recoding auditory representations into a visual format. Experiment 3 precluded the application of any

visual recoding process by equating the thresholds for visual and auditory cues. This equalization reduced any difference in navigation performance between modalities. The stimulus presentation was virtually identical to that used in Experiment 2, except that articulatory suppression was implemented during the spatial navigation task to rule out use of a verbal strategy. Consistent with Experiment 2, Experiment 3 compared single visual, single auditory, and mixture conditions. We examined whether there were any differences between the 1) mixed modality and single conditions, and the 2) single auditory and single visual conditions. No performance differences were expected between the single and mixed modalities, based on the notion that a spatial updating system relies on a non-modality-dependent system in working memory. Lower performance in the auditory modality condition relative to the visual modality condition would indicate a cost of recoding auditory cues to a visual format.

Method

Participants. A new group of 30 students (15 females; mean age = 22.0 years; range: 19–31 years) was recruited for this experiment. Another group of three participants (one female participant; all participants were aged 20 years) did not complete the experimental trials and thus were excluded from the analysis.

Apparatus and stimuli. The apparatus and stimuli were the same as those used in

Experiment 2A, except that the arrow stimuli were replaced with coherent motion displays (random dots as provide a visual directional cue; Figure 7). A motion display containing 100 white dots ($0.03^\circ \times 0.03^\circ$ each) was displayed inside a red rectangle subtending $2.8^\circ \times 2.8^\circ$ at the center of the computer monitor, with a black background. Up to 100% of the dots moved in the same direction, according to the coherence parameter with dot velocity of $21.9^\circ/\text{s}$. Coherence was determined using the PEST algorithm for each participant (see below for details).

Procedure and statistical analysis. Participants completed, in order, a hearing pretest, a pretest for establishing coherence, and a mental-pathway task. The procedure for the hearing pretest was the same as in Experiment 1. The mean accuracy for the 30 participants across the four directions was 94.79% ($SD = 0.06$).

As stated above, the coherence of the visual cue was determined using the PEST algorithm (see Hall, 1974; Taylor & Creelman, 1967) to equalize the discrimination thresholds of the hearing pretest across the four directions for each participant. During the pretest, each participant indicated the direction (right, left, up, or down) of the random dots by pressing the 'd', 'a', 'w' or 's' keys, respectively. The initial coherence was 40% and this increased in steps of 16%. If the hearing pretest accuracy was below 100%, the run ended when the step size decreased to 1% coherence. If the hearing

pretest accuracy was 100% and there were no error responses, the run ended after 30 trials. If there were error responses during the 30 trials, the coherence was increased according to the step size and continued for another 30 trials. Before undertaking the pretest, participants were familiarized with the task through 12 practice trials. The mean coherence of the 30 participants was 27.77% ($SD = 10.54$) across the four directions.

Figure 7 shows a schematic diagram of a single trial of the mental-pathway task. The procedure was virtually identical to that used in Experiment 2A, except for the use of the coherent motion displays as visual directional cues, and for the requirement to utter 'za' twice per second. The modality of the directional cues was either visual, auditory, or mixed and was manipulated blockwise, with the order counterbalanced across participants. The sequence lengths of the directional cues (six or eight) were pseudo-randomly manipulated within each block. Altogether, participants completed a total of 72 trials consisting of three blocks with 24 trials each, yielding 12 trials for each condition of modality \times sequence length of within-subject factors. All participants performed two practice trials prior to the 72 test trials. The percentage of correct responses was calculated for each experimental factor per participant. An ANOVA was conducted with the mean percentage of correct responses as within-subject factors of modality (visual, auditory, or mixed), and sequence length (six or eight).

We analyzed the error patterns with respect to the final choice of target location across the different modality conditions. The aim was to determine whether the matrix cell chosen by participants misidentifying the final target location was selected in a random or systematic manner. A cell in a previously occupied location might be expected to be identified more frequently. This analysis was expected to provide insight into the degree of similarity between visuo- and audio-spatial working memory. Similar error-choice patterns across modalities would support the view that working memory does not differ by modality. Experiment 3 focused on three cells in the target pathway (Figure 9, right), namely the second cell from the beginning of the sequence, the middle cell, and the penultimate cell. To assess participants' cell-choice behavior, we calculated the rates of choosing each cell position, by dividing the number of trials in which participants selected the second, middle, and penultimate cells by the number of all error trials, respectively.

Results and Discussion

It is possible that a long response duration would allow verbal labeling of the information. To omit this possibility, responses (5.51% of all trials) that deviated by more than 1.5 times the interquartile range (IQR) were excluded from the ANOVA. Figure 8 displays the mean percentage of correct responses as a function of sequence

length (six or eight) for the three modalities (visual, auditory, or mixed). The result revealed a main effect of sequence length, $BF_{10} = 56.64$. Importantly, the effect of modality favored the null hypothesis, $BF_{10} = 0.32$. The interaction term supported the null hypothesis, $BF_{10} = 0.15$. Post-hoc tests revealed no differences in the mean percentage of correct responses between the visual and auditory modalities, $BF_{10} = 0.24$, or between the auditory modality and mixed modalities, $BF_{10} = 0.24$. The alternative hypothesis was not conclusively supported by the results of the comparison between the visual and mixed modalities, $BF_{10} = 1.12$. Performance was generally lower in Experiment 3 than in Experiment 2A, likely due to the dual-task condition that was induced by articulatory suppression.

We obtained a total of 1184 error responses (54.81% of all trials) across modality conditions. The error-pattern analysis revealed that 38.00% of the cells reported as erroneous responses were on the target pathway. Random cell choice would be reflected in a rate of choosing cells on the pathway of 4.96% for a sequence length of six (= 6/121 cells) and 6.61% for a sequence length of eight (= 8/121 cells). A Bayesian one-sample *t*-test indicated that the rate of 38.00% was above these chance levels, $BF_{10} = 1.06 \times 10^{25}$, indicating that participants chose cells in a systematic manner.

Figure 9 (left) displays the mean rates of choosing the second, middle, and

penultimate cell positions within the target pathway. Because the main effect of sequence length (six or eight) on cell choice favored the null hypothesis, $BF_{10} = 0.17$, we pooled the data across the sequence lengths. We conducted a two-way ANOVA with modality (visual, auditory, or mixed) and cell position (second, middle, or penultimate) as within-subject factors. There was a main effect of cell position, $BF_{10} = 1.78 \times 10^6$; however, the main effect of modality did not conclusively support the null hypothesis, $BF_{10} = 0.53$. Importantly, the interaction effect supported the null hypothesis, $BF_{10} = 0.06$. Post-hoc tests revealed that the cell-choice rate was greater under the mixed modality than under the auditory modality, $BF_{10} = 6.33$. However, there was no difference between the visual and mixed conditions, $BF_{10} = 0.12$, and the difference between the auditory and visual conditions was inconclusively supported, $BF_{10} = 1.13$. Further post-hoc tests revealed that the penultimate cell was erroneously selected more frequently compared to the second, $BF_{10} = 1.57 \times 10^3$, and middle cells, $BF_{10} = 8.10 \times 10^2$. The second and middle cells were selected at a similar rate, $BF_{10} = 0.13$.

Through application of the PEST procedure, the visual and auditory discrimination thresholds were equalized. In contrast to the findings of Experiment 2, the present results revealed no advantages of the visual modality over the auditory condition with respect to navigation performance. This result implies no cost of transforming auditory

information to the visual domain.

Regarding the results of the mixed modality condition, the BF_{10} analysis showed a trend analogous to that found in Experiments 2A and 2B, in that participants' performances under single-auditory and mixed-modality conditions were comparable. Importantly, there was no modality effect on navigation accuracy. The performances under the visual and mixed modalities were similar, although the BF_{10} analysis did not conclusively favor the null or alternative hypothesis. This result is consistent with the idea of common spatial representation (Lehnert & Zimmer, 2006, 2008b); thus, the current results indicate similarity of visuo- and audio-spatial working memory processes during navigation tasks.

Experiment 3 analyzed patterns of errors in the choice of final target position. Participants more frequently chose the penultimate cell position than the second and middle cells close to the initial target position. The frequency rates of choosing the second and middle cells were similar. This result is consistent with the general consensus that working memory can maintain representations of up to four objects simultaneously (Cowan, 2016; Luck & Vogel, 1997), reflecting that participants successfully remembered the goal location or its vicinity. The results were similar regardless of modality, in accordance with the non-interaction effects of modality \times cell

position. The auditory and mixed conditions were similar in terms of cell choice.

However, the slight differences are intriguing, although it should be noted that this analysis may have been underpowered because it was based on approximately half of all trials (i.e., 54.81%).

To provide further discussion on the interaction effects of modality \times cell position, we depict the total error-choice patterns with respect to final target location in Figure 10. To visualize the error distribution, we aligned individual error responses so that a cell position (i.e., correct answer) was located at the center of the matrix. The results suggest that the error reports were distributed near the correct position irrespective of the modality conditions. Thus, the present results support the idea that error patterns do not differ by modality.

General Discussion

The present study examined the similarity/dissimilarity of visual and auditory spatial working memory functions using an updating task (Cornoldi et al., 1991; Kerr, 1993) across four experiments. The results of Experiment 1 demonstrated a positive correlation between task performance under the visual and auditory modalities. We assessed whether there was a modality-switching cost (e.g., Kreutzfeldt et al., 2017; Loomis et al., 2012; Suchan et al., 2006) as an indicator of a modality-specific influence

on maintaining and updating spatial representations in working memory (e.g., Loomis et al., 2012). The results of Experiments 2A and 2B revealed that modality switching resulted in no performance deficits during working memory navigation. Experiment 3 supported the findings of Experiments 1 and 2. Specifically, participants encoded the spatial information into a common format of representation and used this to update the target's route during the maneuvers. This assumption is consistent with the findings of a neuroimaging study (Lehnert & Zimmer, 2008a) that reported brain activity reflecting common encoding of visual and auditory information while maintaining an object's locations. Experiment 3 ruled out effects of visual or verbal recoding, and of baseline differences across modalities, on the results of Experiments 1 and 2.

The cell choice patterns (whether participants misidentify a target destination in a random or systematic manner during the working memory navigation task) depend on whether spatial working memory is modality dependent. Specifically, if the updated contents are processed in working memory in a similar manner for different modalities, the error patterns of the navigation task should be similar. Alternatively, if working memory updating is modality-dependent, error patterns should differ by modality. The results demonstrated that participants misidentified the target destination in a systematic manner, as reflected by the similar error patterns among the visual, auditory, and mixed

conditions. This finding strongly suggests that spatial updating relies on a non-modality-dependent system, rather than a modality-dependent system.

Experiment 1 demonstrated that performance tended to decrease as the sequence length increased, according to the capacity limitations of working memory (e.g., Luck & Vogel, 1997). The limited capacity of working memory may have affected error-response patterns such that participants frequently chose cells in the vicinity of the target positions in Experiment 3. This error pattern implies that, when the participants were no longer able to follow the target's trajectory, they tended to respond to the position they remembered, which was in the vicinity of the target destination maintained in working memory, rather than that at the beginning of the sequence. Taken together, the results indicate that spatial updates during the matrix task relied on a non-modality-dependent system associated with maintenance processes (Martinkauppi et al., 2000; Lehnert & Zimmer, 2006). The capacity for maintaining spatial contents is limited to post-maintenance processing, such as spatial updating.

Although there is a clear functional and neurophysiological distinction between the working-memory functions of maintenance (e.g., Vogel et al., 2005) and updating (e.g., Bledowski et al., 2004; Polich, 2007), the present study found a trend analogous to that seen in studies on the maintenance function of working memory (Martinkauppi et al.,

2000; Lehnert & Zimmer, 2006, 2008a) in that vision and audition shared the same component. We suggest that this is because updating employs the maintenance process to manipulate and integrate spatial contents in memory. In other words, updating spatial representation would be faster and more accurate when the relevant system is no longer modality-specific but domain-specificity (i.e., spatial domain) contributes to the behavior (e.g., Loomis et al., 2012). We believe that domain-specific results, based on maintenance and updated processing, are informative findings for a working-memory study because the likelihood of modality specificity may differ by working memory stage (e.g., among encoding, retention, and retrieval; Lehnert & Zimmer, 2008b).

The main finding of the present study is consistent with the theoretical position of shared spatial working memory. The modality non-specific view is based on supramodal storage (Martinkauppi et al., 2000; Lehnert & Zimmer, 2006, 2008b), and posits a unitary storage system for working memory or the coexistence of inputs from different modalities in the same component (Lehnert & Zimmer, 2006). Within this view, spatial information may be maintained in the episodic buffer component (Baddeley, 2002; Baddeley et al., 2011) in the model of tripartite working memory (Cowan, 1999; Hitch et al., 2020). Lehnert and Zimmer suggested that a key role for information maintenance is binding objects to location information. Accordingly, spatial information would be

linked to an object's features such as color, shape, or pitch as episodic traces (Lehnert & Zimmer, 2006). Although there is modality-specific dissociation in information streams, this may contribute to representations of the appearance of objects but be irrelevant to the maintenance function (e.g., Lehnert & Zimmer, 2008a).

Despite these findings, how visual recoding is achieved during the processing stages of working memory remains unclear. The results for Experiment 3 ruled out any involvement of visual recoding in updating tasks using visual and auditory cues with equal thresholds. However, previous studies have implied that recoding might occur during maintenance processes, or at an earlier stage (Lehnert & Zimmer, 2008a), to integrate auditory stimuli into visual representations. Visual recoding depends on whether this integration process is required for a given task. Visual representations may have been present in the study of Lehnert and Zimmer (2008a), because their auditory stimuli were sounds (e.g., barking) corresponding to pictures of objects (e.g., a dog). However, the present study cannot be interpreted in the same way, because our matrix task required spatial processes to retain and retrieve the location-related matrix configurations for the different modalities, rather than object-related configurations (e.g., the shape of the visual stimulus or the pitch of the auditory stimulus). The putative system in the present study is more concerned with dynamic information, such as the

movement paths of stimuli (e.g., Logie, 1995, 2011), than with the static information associated with bound objects (Lehnert and Zimmer, 2008a).

Taken together, the current findings support the theoretical position that vision and audition share a component of working memory for spatial processes. This study emphasized multimodal aspects of the updating function of working memory by applying this type of methodology to evaluate the possibility of a modality-switching cost. Performance did not decline when input modalities were switched. In sharing the same system across sensory inputs, working memory may contribute to maintaining and updating spatial representations accurately.

Acknowledgement

This work was supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (17H02648 and 20H01779) to JK and Graduate Grant Program of Graduate School of Letters, Hokkaido University, Japan, and a Grant-in-Aid from the Japan Society for the Promotion of Science Fellows (20J20490) to TM.

Open practices statement

The data and significant program code will be made available after acceptance via the Open Science Framework (<https://osf.io/sf742/>), and none of the experiments were preregistered.

Reference

- Alais, D., & Burr, D. (2004). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology*, *14*(3), 257–262.
[https://doi.org/10.1016/S0960-9822\(04\)00043-0](https://doi.org/10.1016/S0960-9822(04)00043-0)
- Alain, C., He, Y., & Grady, C. (2008). The contribution of the inferior parietal lobe to auditory spatial working memory. *Journal of Cognitive Neuroscience*, *20*(2), 285–295. <https://doi.org/10.1162/jocn.2008.20014>
- Alain, C., McDonald, K. L., Kovacevic, N., & McIntosh, A. R. (2009). Spatiotemporal analysis of auditory "what" and "where" working memory. *Cerebral Cortex*, *19*(2), 305–314. <https://doi.org/10.1093/cercor/bhn082>
- Attneave, F., & Curlee, T. E. (1983). Locational representation in imagery: A moving spot task. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(1), 20–30. <https://doi.org/10.1037/0096-1523.9.1.20>
- Awh, E., Jonides, J., & Reuter-Lorenz, P. A. (1998). Rehearsal in spatial working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 780–790. <https://doi.org/10.1037/0096-1523.24.3.780>
- Baddeley, A. D. (2002). Is working memory still working? *European Psychologist*, *7*(2), 85–97. <https://doi.org/10.1027//1016-9040.7.2.85>
- Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29. <https://doi.org/10.1146/annurev-psych-120710-100422>
- Baddeley, A. D., Allen, R. J., & Hitch, G. J. (2011). Binding in visual working memory: The role of the episodic buffer. *Neuropsychologia*, *49*(6), 1393–1400.
<https://doi.org/10.1016/j.neuropsychologia.2010.12.042>

- Bledowski, C., Prvulovic, D., Hoehstetter, K., Scherg, M., Wibrals, M., Goebel, R., & Linden, D. E. J. (2004). Localizing P300 Generators in Visual Target and Distractor Processing: A Combined Event-Related Potential and Functional Magnetic Resonance Imaging Study. *The Journal of Neuroscience*, *24*(42), 9353–9360.
<https://doi.org/10.1523/JNEUROSCI.1897-04.2004>
- Bushara, K. O., Weeks, R. A., Ishii, K., Catalan, M.-J., Tian, B., Rauschecker, J. P., & Hallett, M. (1999). Modality-specific frontal and parietal areas for auditory and visual spatial localization in humans. *Nature Neuroscience*, *2*(8), 759–766.
<https://doi.org/10.1038/11239>
- Craik, F. I. (1969). Modality effects in short-term storage. *Journal of Verbal Learning & Verbal Behavior*, *8*(5), 658–664. [https://doi.org/10.1016/S0022-5371\(69\)80119-2](https://doi.org/10.1016/S0022-5371(69)80119-2)
- Cohen, M. A., Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). Auditory and visual memory in musicians and nonmusicians. *Psychonomic Bulletin & Review*, *18*(3), 586–591. <https://doi.org/10.3758/s13423-011-0074-0>
- Cohen, M. A., Horowitz, T. S., & Wolfe, J. M. (2009). Auditory recognition memory is inferior to visual recognition memory. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *106*(14), 6008–6010.
<https://doi.org/10.1073/pnas.0811884106>
- Colby, C. L., & Goldberg, M. E. (1999). Space and attention in parietal cortex. *Annual Review of Neuroscience*, *22*, 319–349. <https://doi.org/10.1146/annurev.neuro.22.1.319>
- Cornoldi, C., Cortesi, A., & Preti, D. (1991). Individual differences in the capacity limitations of visuospatial short-term memory: Research on sighted and totally congenitally blind people. *Memory & Cognition*, *19*(5), 459–468.
<https://doi.org/10.3758/BF03199569>

- Cornoldi, C., Tinti, C., Mammarella, I. C., Re, A. M., & Varotto, D. (2009). Memory for an imagined pathway and strategy effects in sighted and in totally congenitally blind individuals. *Acta Psychologica, 130*(1), 11–16.
<https://doi.org/10.1016/j.actpsy.2008.09.012>
- Cowan, N. (1999). An Embedded-Processes Model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (p. 62–101). Cambridge University Press.
<https://doi.org/10.1017/CBO9781139174909.006>
- Cowan, N. (2016). *Working Memory Capacity*. New York: Routledge,
<https://doi.org/10.4324/9781315625560>
- Delogu, F., Nijboer, T. C. W., & Postma, A. (2012). Encoding location and serial order in auditory working memory: Evidence for separable processes. *Cognitive Processing, 13*(3), 267–276. <https://doi.org/10.1007/s10339-012-0442-3>
- Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review, 25*(1), 207–218.
<https://doi.org/10.3758/s13423-017-1266-z>
- Driver, J., & Spence, C. (2004). Crossmodal spatial attention: Evidence from human performance. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention*. Oxford: Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780198524861.003.0008>
- Goossens, H. H. L. M., & Van Opstal, A. J. (1999). Influence of head position on the spatial representation of acoustic targets. *Journal of Neurophysiology, 81*(6), 2720–2736.
- Hall, J. L. (1974) PEST: Note on the reduction of variance of threshold estimates. *The*

Journal of the Acoustical Society of America, 55(5), 1090–1091.

<https://doi.org/10.1121/1.1914656>

Hitch, G. J., Allen, R. J., & Baddeley, A. D. (2020). Attention and binding in visual working memory: Two forms of attention and two kinds of buffer storage.

Attention, Perception, & Psychophysics, 82(1), 280–293.

<https://doi.org/10.3758/s13414-019-01837-x>

JASP Team (2019). JASP (Version 0.11.1) [Computer software], Available at

<https://jasp-stats.org/>

Javitt, D. C., Strous, R. D., Grochowski, S., Ritter, W., & Cowan, N. (1997). Impaired precision, but normal retention, of auditory sensory ("echoic") memory information in schizophrenia. *Journal of Abnormal Psychology*, 106(2), 315–324.

<https://doi.org/10.1037/0021-843X.106.2.315>

Jeffreys, H. (1961). *Theory of probability, 3rd edn.* Oxford, UK: Oxford University Press.

Kaiser, J. (2015). Dynamics of auditory working memory. *Frontiers in Psychology*, 6, Article 613.

Kerr, N. H. (1993). Rate of imagery processing in two versus three dimensions.

Memory & Cognition, 21(4), 467–476. <https://doi.org/10.3758/BF03197178>

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., &

Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin*, 136(5), 849–874. <https://doi.org/10.1037/a0019842>

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3?

Perception, 36(14), 1–16, *ECVP Abstract Supplement*.

Kong, L., Michalka, S. W., Rosen, M. L., Sheremata, S. L., Swisher, J. D., Shinn-

- Cunningham, B. G., & Somers, D. C. (2014). Auditory spatial attention representations in the human cerebral cortex. *Cerebral Cortex*, *24*(3), 773–784. <https://doi.org/10.1093/cercor/bhs359>
- Kreutzfeldt, M., Stephan, D. N., Willmes, K., & Koch, I. (2017). Modality-specific preparatory influences on the flexibility of cognitive control in task switching. *Journal of Cognitive Psychology*, *29*(5), 607–617. <https://doi.org/10.1080/20445911.2017.1293064>
- Krumbholz, K., Nobis, E. A., Weatheritt, R. J., & Fink, G. R. (2009). Executive control of spatial attention shifts in the auditory compared to the visual modality. *Human Brain Mapping*, *30*(5), 1457–1469. <https://doi.org/10.1002/hbm.20615>
- Lehnert, G., & Zimmer, H. D. (2006). Auditory and visual spatial working memory. *Memory & Cognition*, *34*(5), 1080–1090. <https://doi.org/10.3758/BF03193254>
- Lehnert, G., & Zimmer, H. D. (2008a). Common coding of auditory and visual spatial information in working memory. *Brain Research*, *1230*, 158–167. <https://doi.org/10.1016/j.brainres.2008.07.005>
- Lehnert, G., & Zimmer, H. D. (2008b). Modality and domain specific components in auditory and visual working memory tasks. *Cognitive Processing*, *9*(1), 53–61. <https://doi.org/10.1007/s10339-007-0187-6>
- Lewald, J., & Ehrenstein, W. H. (2001). Spatial coordinates of human auditory working memory. *Cognitive Brain Research*, *12*(1), 153–159. [https://doi.org/10.1016/S0926-6410\(01\)00042-8](https://doi.org/10.1016/S0926-6410(01)00042-8)
- Logie, R. H. (1995). *Essays in cognitive psychology. Visuo-spatial working memory*. Lawrence Erlbaum Associates, Inc.
- Logie, R. H. (2011). The functional organization and capacity limits of working

- memory. *Current Directions in Psychological Science*, 20(4), 240–245.
<https://doi.org/10.1177/0963721411415340>
- Loomis, J. M., Klatzky, R. L., McHugh, B., & Giudice, N. A. (2012). Spatial working memory for locations specified by vision and audition: Testing the amodality hypothesis. *Attention, Perception, & Psychophysics*, 74(6), 1260–1267.
<https://doi.org/10.3758/s13414-012-0311-2>
- Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657), 279–281. <https://doi.org/10.1038/36846>
- Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, 17(8), 391–400. <https://doi.org/10.1016/j.tics.2013.06.006>
- Martinkauppi, S., Rämä, P., Aronen, H. J., Korvenoja, A., & Carlson, S. (2000). Working memory of auditory localization. *Cerebral Cortex*, 10(9), 889–898.
<https://doi.org/10.1093/cercor/10.9.889>
- Mance, I., & Vogel, E. K. (2013). Visual working memory. *WIREs Cognitive Science*, 4(2), 179–190. <https://doi.org/10.1002/wcs.1219>
- Michalka, S. W., Rosen, M. L., Kong, L., Shinn-Cunningham, B. G., & Somers, D. C. (2015). Auditory spatial coding flexibly recruits anterior, but not posterior, visuotopic parietal cortex. *Cerebral Cortex*, 26(3), 1302–1308.
<https://doi.org/10.1093/cercor/bhv303>
- Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, 7(3), 134–140.
[https://doi.org/10.1016/S1364-6613\(03\)00028-7](https://doi.org/10.1016/S1364-6613(03)00028-7)
- Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(4), 406–419.

<https://doi.org/10.1037/a0024377>

- Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). BayesFactor: Computation of Bayes factors for common designs (R package, version 0.9.12-4.2) [Computer software]. Available at: <https://CRAN.R-project.org/package=BayesFactor>
- Polich, J. (2007). Updating p300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128–2148. <https://doi.org/10.1016/j.clinph.2007.04.019>
- Quak, M., London, R. E., & Talsma, D. (2015). A multisensory perspective of working memory. *Frontiers in Human Neuroscience*, *9*, Article 197.
- Ricker, T. J., AuBuchon, A. M., & Cowan, N. (2010). Working memory. *WIREs Cognitive Science*, *1*(4), 573–585.
- Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, *22*(2), 304–321. <https://doi.org/10.1037/met0000057>
- Ruthruff, E., Remington, R. W., & Johnston, J. C. (2001). Switching between simple cognitive tasks: The interaction of top-down and bottom-up factors. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(6), 1404–1419. <https://doi.org/10.1037/0096-1523.27.6.1404>
- Sinnett, S., Juncadella, M., Rafal, R., Azañón, E., & Soto-Faraco, S. (2007). A dissociation between visual and auditory hemi-inattention: Evidence from temporal order judgements. *Neuropsychologia*, *45*(3), 552–560. <https://doi.org/10.1016/j.neuropsychologia.2006.03.006>
- Smith, E. E., & Jonides, J. (1997). Working memory: A view from neuroimaging. *Cognitive Psychology*, *33*(1), 5–42. <https://doi.org/10.1006/cogp.1997.0658>

- Smyth, M. M., & Scholey, K. A. (1994). Interference in immediate spatial memory. *Memory & Cognition*, 22(1), 1–13. <https://doi.org/10.3758/BF03202756>
- Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74(11), 1–29. <https://doi.org/10.1037/h00937593>
- Spence, C., Nicholls, M. R., & Driver, J. (2001). The cost of expecting events in the wrong sensory modality. *Perception & Psychophysics*, 63(2), 330–336. <https://doi.org/10.3758/BF03194473>
- Suchan, B., Linnewerth, B., Köster, O., Daum, I., & Schmid, G. (2006). Cross-modal processing in auditory and visual working memory. *Neuroimage*, 29(3), 853–858. <https://doi.org/10.1016/j.neuroimage.2005.08.014>
- Tao, Q., Chan, C. C. H., Luo, Y.-j., Li, J.-j., Ting, K.-h., Lu, Z.-l., Whitfield-Gabrieli, S., Wang, J., & Lee, T. M. C. (2017). Prior visual experience modulates learning of sound localization among blind individuals. *Brain Topography*, 30(3), 364–379. <https://doi.org/10.1007/s10548-017-0549-z>
- Taylor, M., & Creelman, C. D. (1967). PEST: Efficient estimates on probability functions. *The Journal of the Acoustical Society of America*, 41(4A), 782–787. <https://doi.org/10.1121/1.1910407>
- Tremblay, S., Parmentier, F. B. R., Guérard, K., Nicholls, A. P., & Jones, D. M. (2006). A spatial modality effect in serial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5), 1208–1215. <https://doi.org/10.1037/0278-7393.32.5.1208>
- Vogel, E. K., McCullough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, 438(7067),

500–503. <https://doi.org/10.1038/nature04171>

Vuontela, V., Steenari, M.-R., Carlson, S., Koivisto, J., Fjällberg, M., & Aronen, E. T.

(2003). Audiospatial and visuospatial working memory in 6-13 year old school children. *Learning & Memory*, *10*(1), 74–81. <https://doi.org/10.1101/lm.53503>

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... & Morey,

R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. <https://doi.org/10.3758/s13423-017-1323-7>

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045.

<https://doi.org/10.1037/xge0000014>

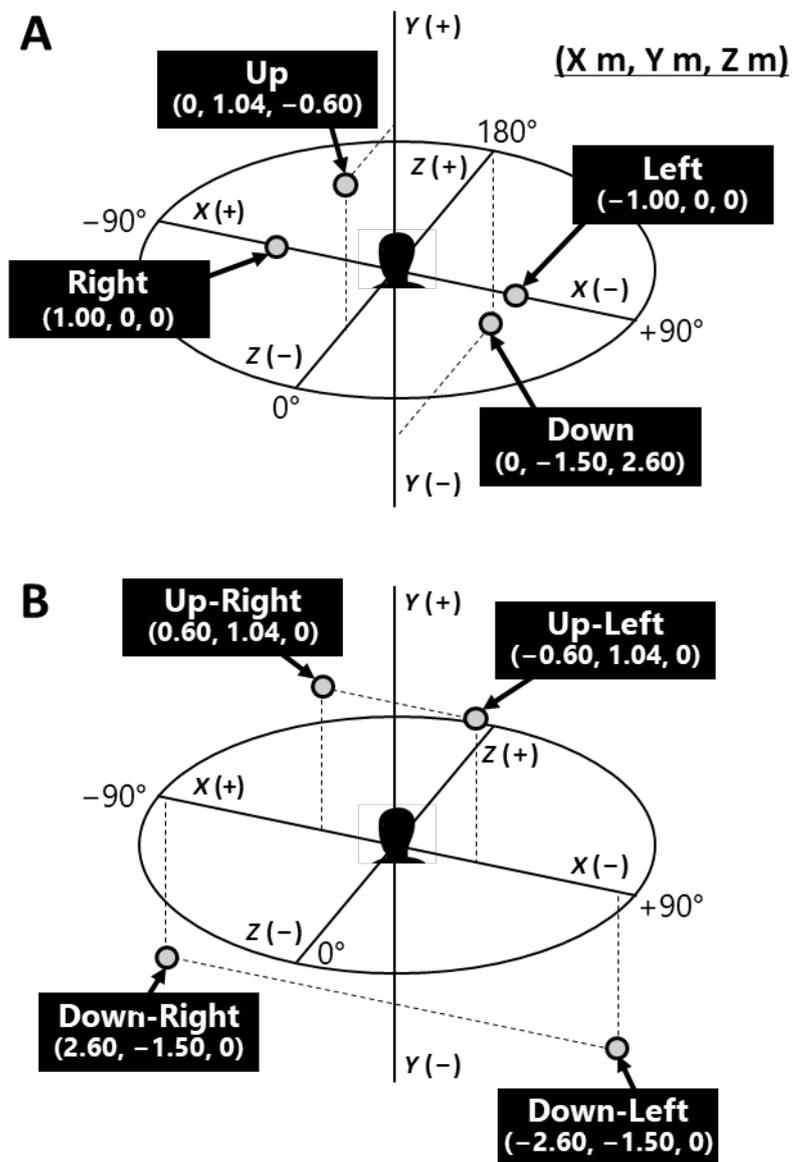
Wetzels, R., Grasman, R. P., & Wagenmakers, E. J. (2012). A default Bayesian

hypothesis test for ANOVA designs. *The American Statistician*, *66*(2), 104–111.

<https://doi.org/10.1080/00031305.2012.695956>

Figure 1.

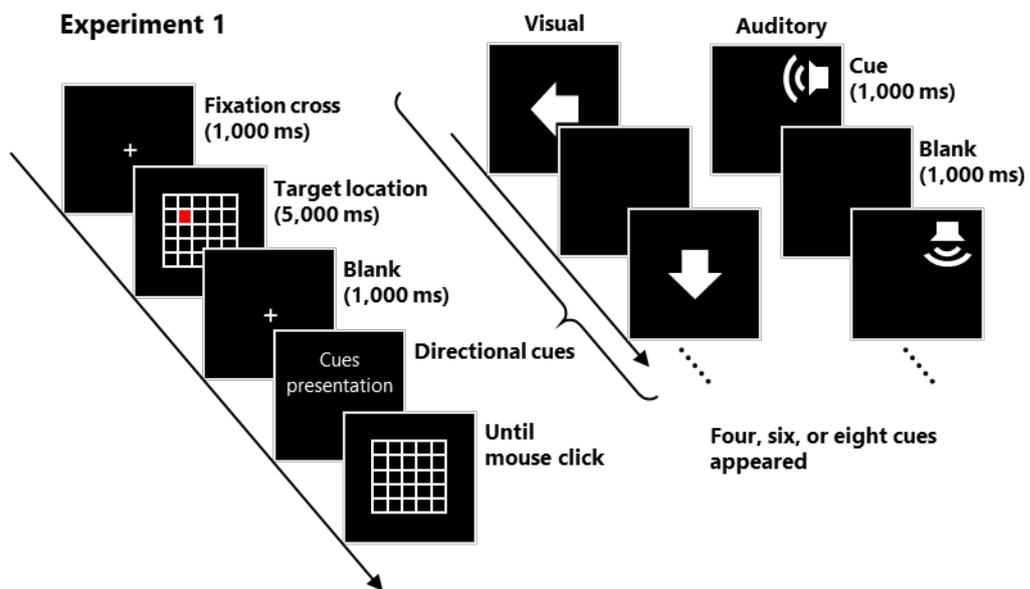
Arrangement of the virtual sound sources relative to a participant's head position



Note. Panel A: the sound source positions in Experiments 1, 2A, and 3. Panel B: the sound source positions in Experiment 2B.

Figure 2.

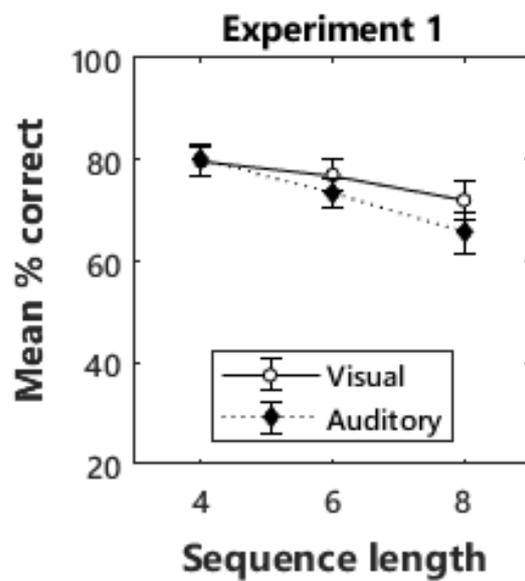
Schematic example of a trial sequence in Experiment 1



Note. The icons represent presentation of a directional auditory cue. No visual stimuli were presented on the screen during presentation of the auditory directional cues.

Figure 3.

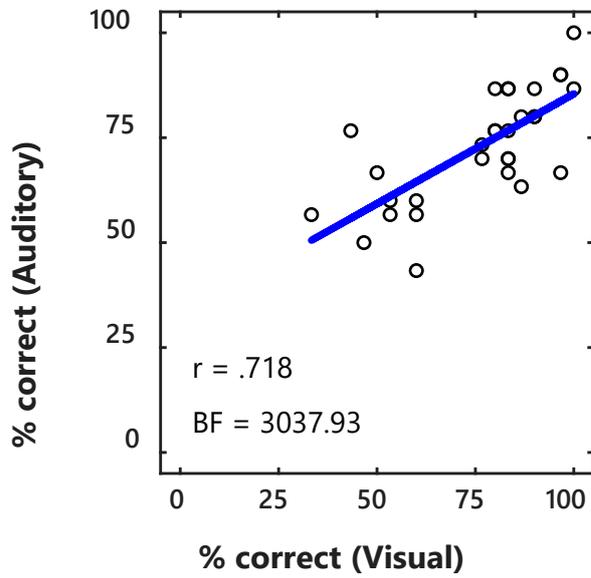
Mean percentages of correct responses in Experiment 1



Note. The mean percentages of correct responses for each modality (visual and auditory) are shown as a function of sequence length (four, six, and eight). Error bars indicate standard error.

Figure 4.

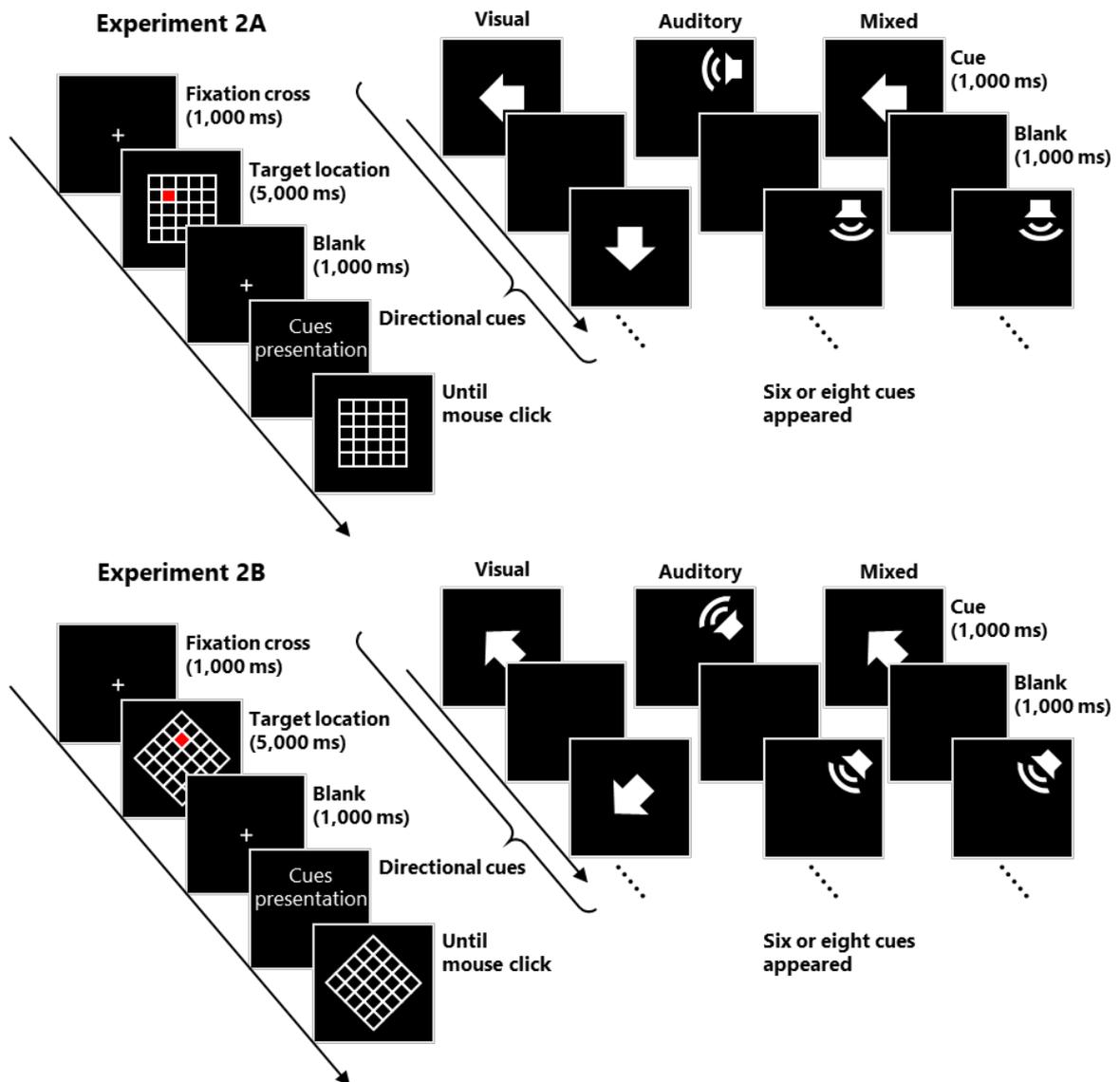
Correlation plot of the percentages of correct responses in Experiment 1



Note. The percentages of correct responses to two modalities (visual and auditory) were averaged across sequence length (four, six, and eight). Several points of the scatter plots overlapped each other.

Figure 5.

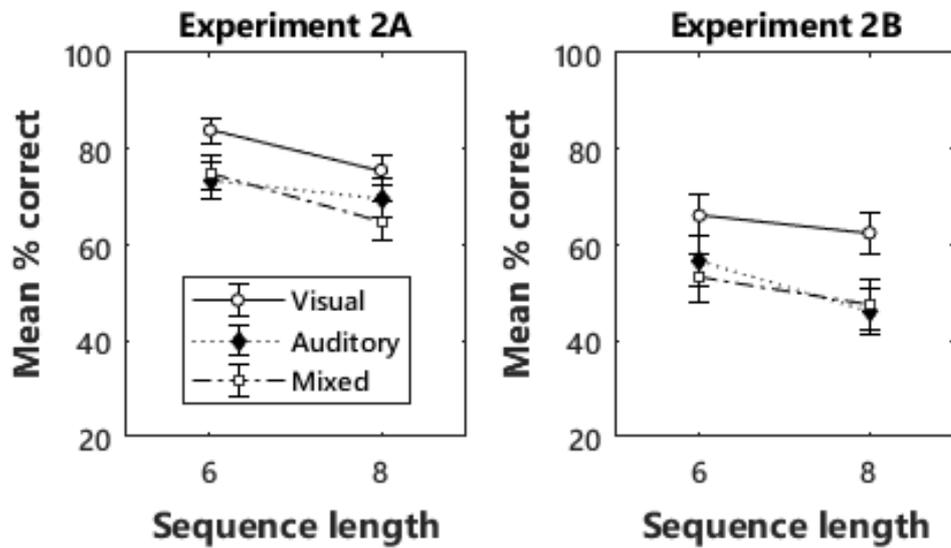
Schematic examples of a trial sequence in Experiment 2A and 2B



Note. No visual stimuli were presented on the screen during presentation of the auditory directional cues.

Figure 6.

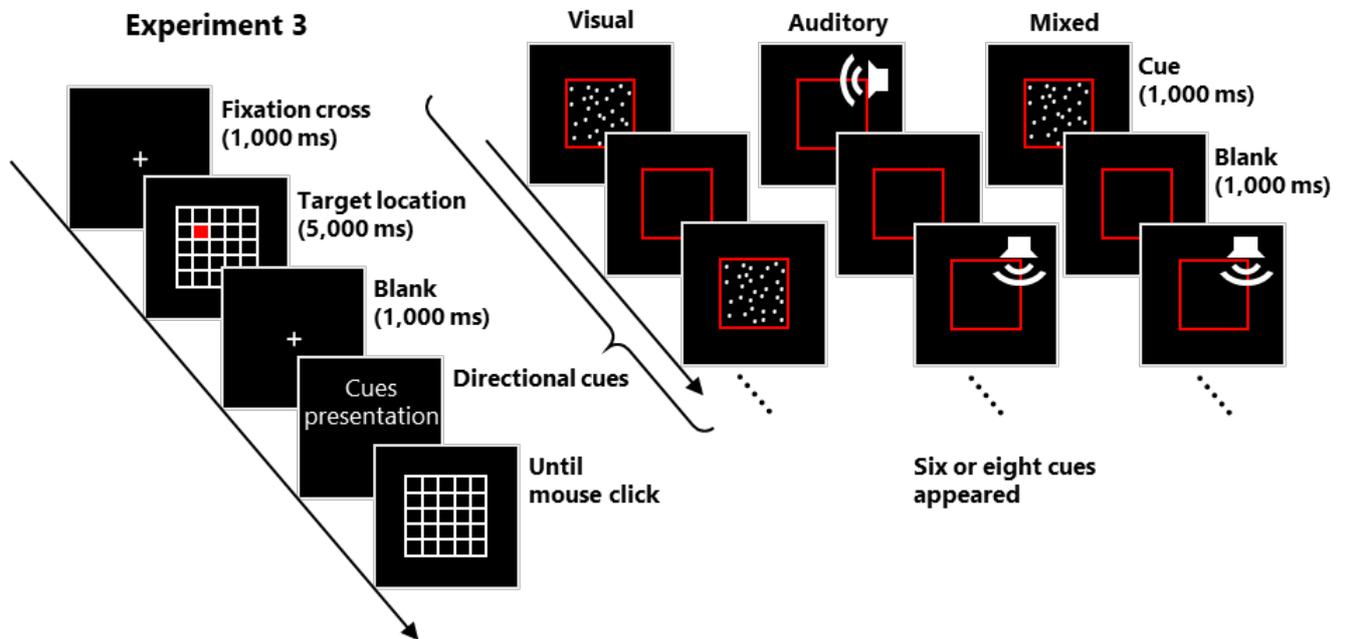
Mean percentages of correct responses in Experiment 2



Note. The mean percentages of correct responses for each modality (visual, auditory, and mixed) are shown as a function of sequence length (six and eight). Error bars indicate standard error.

Figure 7.

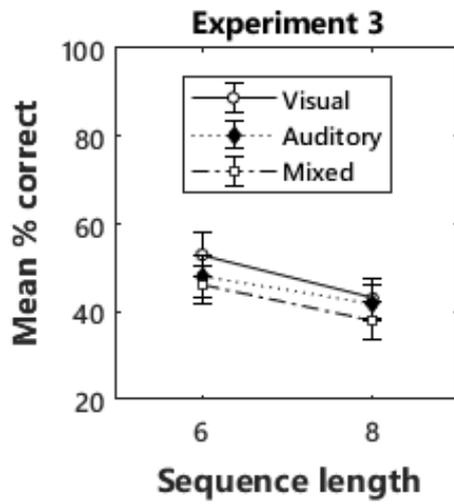
Schematic example of a trial sequence in Experiment 3



Note. No visual stimuli were presented on the screen except for the red-colored rectangular frame during presentation of the auditory directional cues.

Figure 8.

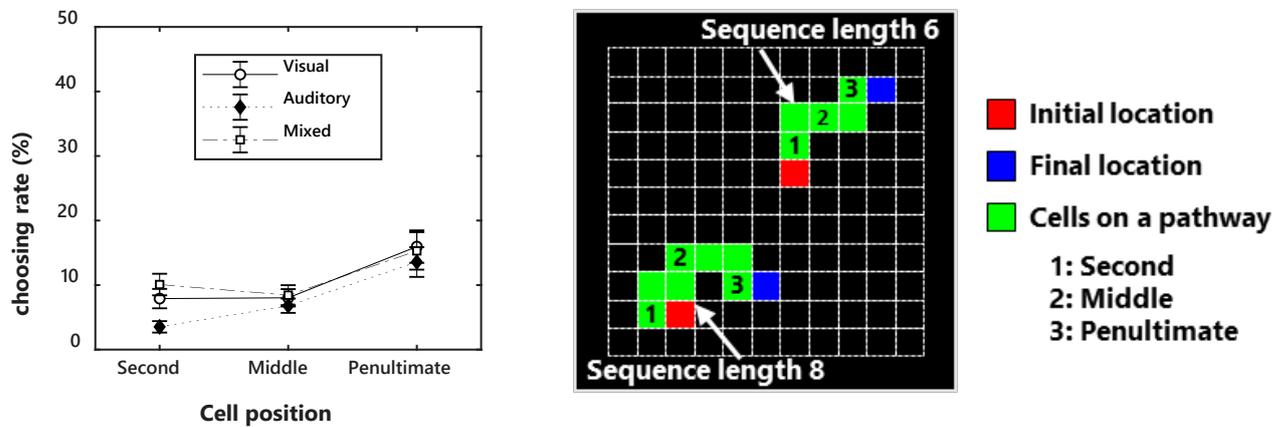
Mean percentages of correct responses in Experiments 3



Note. The mean percentages of correct responses for each modality (visual, auditory, and mixed) are shown as a function of sequence length (six or eight). Error bars indicate standard error.

Figure 9.

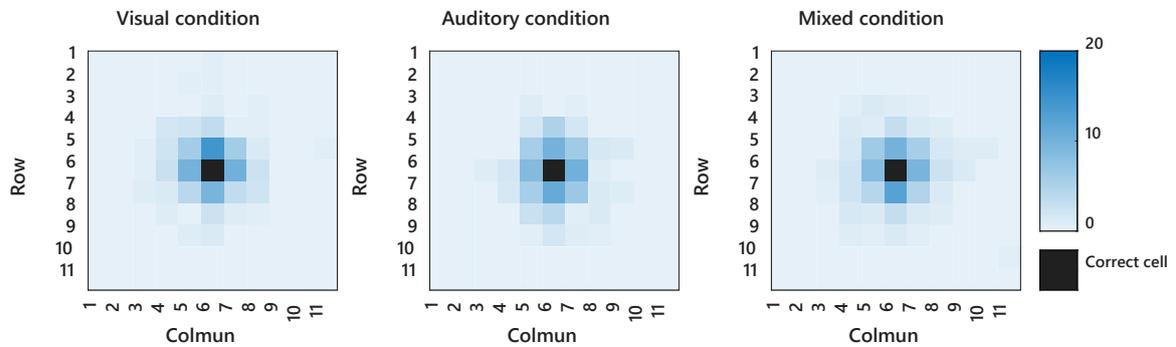
Mean percentages of erroneous cell location choices in Experiment 3



Note. Note. (a) Mean rates of choosing cell position (second, middle, or penultimate) are shown by modality (visual, auditory, or mixed). Error bars indicate standard error. (b) Schematic of a target pathway.

Figure 10.

Patterns of cell choice errors with respect to the final target location



Note. Patterns of cell choice errors with respect to the final target location are shown, with the correct cell position aligned in the center of the matrix. The color bar indicates the frequency of choosing cell position (%). Choices that were not within the matrix (0.6% of all choices) were excluded from the analysis.