Perceptual Load Modulates the Effect of Lightness/Pitch Correspondence on Visual Working Memory Performance

Agnes Rigo[1,2], Hettie Roebuck[1], Barbara Manini[1]

[1]School of Psychology, Department of Health, Psychology and Social Care, University of Derby, Derby, UK

[2]Division of Science and Mathematics, New York University Abu Dhabi, Abu Dhabi, UAE

Data, analysis, and additional online materials are openly available at the project's Open Science Framework page (https://osf.io/m8t5b/). The author(s) declare no potential competing interests regarding the research, authorship, and/or publication of this article. This work was not supported by any funding. This study received ethics approval from the University of Derby.

This paper has not been peer reviewed and it is not the copy of record and may not exactly replicate the authoritative document published in the APA journal.

Corresponding author:
Hettie Roebuck, School of Psychology, Department of Health, Psychology and Social Care, University of Derby, Derby, United Kingdom, h.roebuck@derby.ac.uk

Abstract

Prior studies demonstrate that cross-modal correspondence, the seemingly arbitrary association of features across different sensory modalities, enhances working memory performance. Incongruent background noise has also been shown to aid visual working memory. The current study builds on these premises to investigate the effect of lightness/pitch audiovisual correspondence on visual working memory. We designed a black-and-white orientation change detection task, where the visual stimuli were paired with high- and low-pitched sounds. We compared change-detection performance in conditions with audiovisual correspondence with performance under non-corresponding and visual-only conditions. Additionally, we explored the impact of perceptual load, the stage of memory processing during which audiovisual correspondence is displayed, and the direction of attention to the auditory modality. We found that, in the lightness/pitch domain, cross-modal correspondence does not automatically enhance visual working memory accuracy and reaction time; instead, the salience of the effect is moderated by perceptual load. Lightness/pitch correspondence improved performance only under high perceptual load, with the effect being strongest when corresponding stimuli were displayed during both memory encoding and recall. In contrast, under conditions of low perceptual load, the mere presence of auditory pitch, irrespective of cross-modal correspondence, improved performance, likely by increasing alertness. The study demonstrates that lightness/pitch correspondence does not depend on conscious selective attention and suggests that cross-modal correspondence may serve a functional role beyond sensory integration. We frame our interpretation considering

Bayesian theory with the cognitive system relying more heavily on statistical learning principles under high perceptual demands.

Key words: visual working memory; cross-modal correspondence; audiovisual stimuli; perceptual load; change detection task; Bayesian learning.

Public significance statement:

- Our results suggest that lightness/pitch cross-modal correspondence is automatic and not influenced by conscious attentional focus to simultaneous auditory input.
- This study proposes that the brain utilizes cross-modal correspondence as an information processing strategy, influenced by principles of statistical learning, when faced with high perceptual load. Under conditions of low perceptual load, the mere presence of auditory pitch alongside visual stimuli enhanced both accuracy and reaction time in an orientation change detection task. However, under high perceptual load, only the audiovisual cross-modally corresponding conditions had an enhancing effect.
- These findings contribute to the understanding of how the brain optimizes information processing across senses. The findings may inform strategies for enhancing information processing.

Introduction

In our daily lives, we experience a constant flow of information. Perception is formed through the integration of simultaneous multisensory stimuli, enabling people to perceive the world as a seamless, unified experience. To achieve coherent cognition and perception, the human brain identifies whether signals received through multiple sensory modalities originate from a common source and unites or segregates them accordingly (Ernst & Bülthoff, 2004; Noppeney, 2021). Cross-modal correspondence (CC) refers to the association of often seemingly arbitrary features across senses. For example, the association between a dark or light color and a high- or low-pitched sound. This correspondence has been shown to aid the integration and interpretation of multisensory information, contributing to coherent perception of the environment (Spence, 2011).

Prior research indicates that audiovisual intersensory binding is influenced by several types of CCs, such as associations between shape, elevation and darkness in the visual domain and pitch of sound in the auditory domain (Brunetti et al., 2017; Ćwiek et al., 2022). Over the years, the effect of CCs on perception and performance has been investigated in a variety of behavioral tasks (for reviews, see Spence, 2011; Spence & Deroy, 2013). One of the most extensively studied auditory attributes in CC research is pitch (Spence, 2020; Uno, 2022), and it has repeatedly been shown to correspond with visual features such as brightness and darkness (Marks, 1987; McEwan et al., 2024). Research by Zeljko et al. (2019) indicates that light and dark objects are significantly easier to distinguish when they are paired with corresponding high or low pitches. In a

later study, Zeljko et al. (2021) found that lightness/pitch congruence can even influence the perception of ambiguous illusions such as the Rubin (1915) face/vase.

CC not only influences perception but also higher-level cognitive processes such as working memory (WM). For example, research on grapheme-color synesthesia has shown that CC improves WM performance (Terhune et al., 2013). Moreover, Makovac et al. (2014) found that a short delay period (150 ms) between cue and probe, compared to a long one (1150 ms), facilitates the recognition of shapes by a congruent sound and enhances visual working memory (VWM) performance. These findings suggest that the benefits of cross-modal cueing are time-sensitive. In their study on the effect of CC on WM, Brunetti et al. (2017) found that CCs can enhance both WM accuracy and reaction time (RT). The authors also explored the influence of attended modality on the interaction between CC and WM, finding that performance remained stable when participants were instructed to pay attention to audiovisual stimuli over unimodal stimuli.

As we experience a continuous flow of information, some relevant and some irrelevant to our goals, our WM capacity is limited by our finite selective attention. In this context, perceptual load (e.g., the amount of sensory information that needs to be processed) moderates selective attention. Specifically, previous research suggests that a higher perceptual load facilitates filtering out task-irrelevant information (Lavie & Tsal, 1994; Lavie, 1995, Lavie et al., 2004). In the context of cross-modal correspondence, more research is needed to explore how factors that influence perceptual and post-perceptual processes may affect the relationship between CC and WM. The current study aims to fill this gap by investigating the possible modulating effect of perceptual

load on the interaction between CC and VWM. This can provide insight into how attentional mechanisms operate under varying levels of sensory processing demand. Additionally, how perceptual load may, in turn, affect cognitive control and the brain's information processing strategy.

The present study was designed to provide a comprehensive understanding of the effect of lightness/pitch CC on VWM performance. Specifically, we want to explore how the processing stage at which CC is displayed, perceptual load, and modality-specific selective attention affect this interaction. Prior studies measuring the effect of CC on WM have used n-back tasks (e.g., Brunetti et al., 2017; Terhune et al., 2013). While n-back is a popular method for assessing WM processing, it is known to evoke a relatively high cognitive load (Lamichhane et al., 2020). The present study manipulates the amount of perceptual load, utilizing an orientation change detection task designed explicitly for this experiment. Unlike n-back tasks, the change detection paradigm is thought to isolate VWM without extensively involving other perceptual and cognitive processes, which is suggested to result in purer VWM measurements (Kane et al., 2007). Moreover, by using an orientation change detection task perceptual load can be controlled more effectively (Jaeggi et al., 2008).

Past studies indicate that enhancement to VWM is not limited to audiovisual CC, but has also been observed with other sounds, such as white noise and background speech (Han et al., 2013; Han et al., 2021). Therefore, we aimed to investigate the effect of both corresponding and non-corresponding audiovisual stimuli on VWM performance. To assess at which stage of WM cross-modal sensory information might facilitate accuracy and RT, we created four audiovisual conditions by manipulating the

presence and absence of CC during memory encoding (when the sample array was presented) and memory recall (when the target array was presented). We compared these conditions with VWM performance without any audio presentation ('Visual-only stimuli'). To control for perceptual load, we created a 'Low load' and a 'High load' condition of the task by manipulating the number of objects displayed in each array.

In line with previous research that found that audiovisual CC can enhance WM (Brunetti et al., 2017; Makovac et al., 2014), we hypothesized that audiovisual CC would increase participants' performance in the orientation change-detection task, both in terms of accuracy and RT. Specifically, we predicted that performance would be higher during the audiovisual correspondent conditions than during the non-correspondent, or 'Visual-only' conditions. We also assessed if CC has a more significant impact on VWM when it is presented during memory encoding (e.g., at the presentation of the sample stimulus), recall (e.g., at the presentation of the target stimulus), or both stages. If correspondence during encoding improves VWM performance more than in other conditions, it could indicate that the effect of CC on VWM is already present during the early processing stages (Murray et al., 2008; Zlejko et al., 2019) implying that audiovisual correspondence plays a significant role in enhancing the initial formation of memories. However, if CC presented during the memory recall phase increases VWM performance, it might point towards a more post-perceptual effect, supporting the notion that WM selection relies on internally directed shifts of attention that highlight task-relevant information (Zhou et al., 2022). Additionally, if VWM performance is most significantly enhanced in the scenario where CC occurs at both stages, it could mean

that multisensory integration plays a role throughout the entire VWM process in a distributed nature (Christophel et al., 2017; McEwan et al., 2024).

The load theory of Lavie et al. (2004) suggests that individuals have finite attentional capacity, and whether conceptually irrelevant information is processed depends on the perceptual load and attentional demand. Due to limited attentional resources, a high perceptual load prevents distractor processing through early selection (Lavie, 2005). In light of this theory, we predicted that different perceptual loads during the task would have a role in modulating the effect of CC on VWM processing. We hypothesized that CC would have a stronger effect on VWM performance under high perceptual load, where attentional resources are depleted by relevant tasks and non-corresponding auditory stimuli can be more effectively ignored.

Research indicates that the effect of CC on WM remains stable when attending to different modalities (Brunetti et al., 2017). Furthermore, while the attended modality (visual vs. auditory) can have a significant effect on WM performance, paying attention to audiovisual stimuli over unimodal stimuli was not found to have an effect. This aligns with findings of multiple studies which show that observers tend to prioritize signals from the modality considered most informative for the task (e.g., Burr et al., 2009; Butler et al., 2010: Fetsch et al., 2009). However, further research found that modality-specific selective attention reduces multisensory integration (Badde et al., 2020; Mozolic et al., 2008). Therefore, paying attention to both auditory and visual modalities should promote CC. To alleviate this discrepancy, we measured the role of the attended modality (visual vs. audiovisual) in this process. Building upon the studies mentioned above we hypothesized that the attended modality would significantly impact VWM performance.

Specifically, we predicted that participants instructed to attend to both visual and auditory stimuli would perform better than those instructed to attend only to visual stimuli.

Methods

Participants

Participants were recruited online, using the University of Derby's recruitment space and opportunity sampling. While classically VWM research has been conducted in laboratory settings, Ross-Sheehy et al. (2021) found that unsupervised online assessment of VWM yields similar results to laboratory experiments if environmental factors are controlled for.

Fifty-two individuals completed the online study. To determine the target sample size, a priori power analysis was conducted using G*Power (version 3.1.9.7; Faul et al., 2007). Based on a medium effect size (f2 = .25) and using a standard alpha level of .05, a minimum of 22 participants were required to have 80% power in the analyses. We excluded participants who had > 5% missing data and those with low global accuracy < 55% separately for the 'Low load' and 'High load' conditions. Furthermore, we removed participants whose RT fell outside the interquartile range (IQR) of 0.5-2.6 seconds in more than 20% of all trials. After removing outliers, N = 43 participants (Mage = 33.7 years; SD = 10.6; 65.1% female; range = 19-63 years) were included in the analysis for the 'Low load' condition. For the 'High load' condition, N = 41 participants (Mage = 32.2 years; SD = 9.6; 65.9% female; range = 19-59 years) were included in the analysis. This study followed the ethical standards of the British Psychological Society (2021), the American Psychological Association (2017), and was

approved by the University of Derby's research ethics committee. All participants provided informed consent. There was no monetary compensation offered, but those who signed up for the study via the university's research participation scheme received points for participating upon completion. This study was not preregistered.

Transparency and Openness

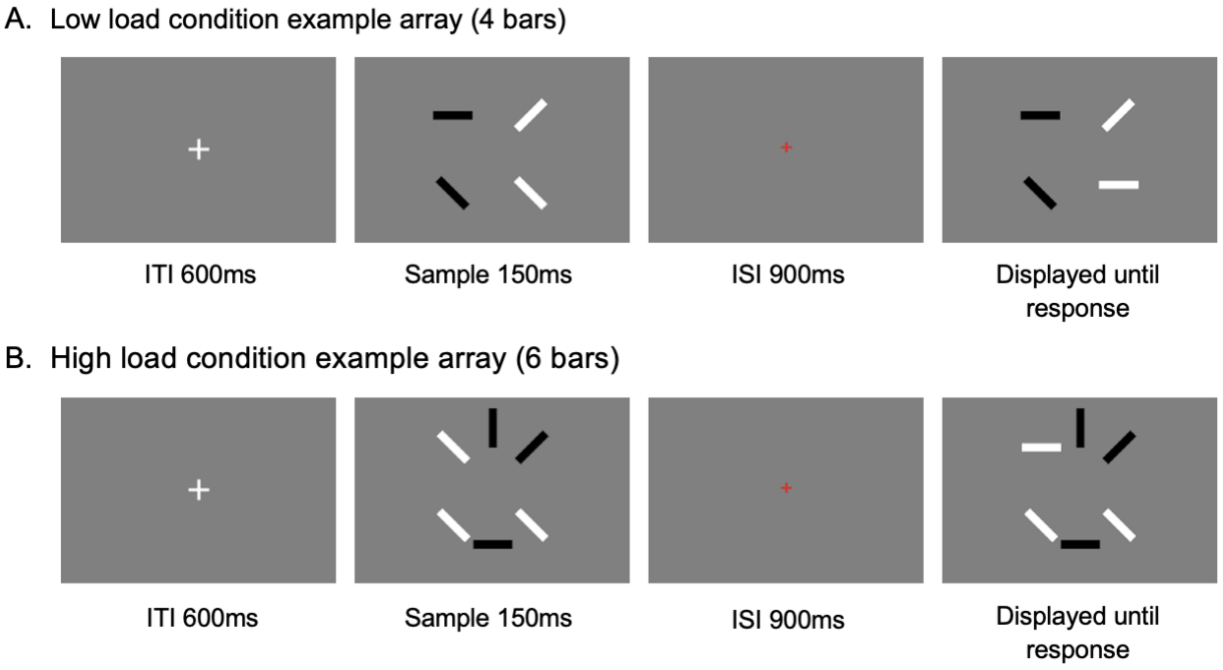
We reported how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The study follows the Journal Article Reporting Standards (Appelbaum, et al., 2018). All data, analysis, and research materials are available at https://osf.io/m8t5b/. The data was analyzed using SPSS version 29.

Apparatus and Stimuli

A black-and-white adaptation of the canonical orientation change detection task (Luck & Vogel, 1997) was built for this study using Pavlovia (https://pavlovia.org). In this task, participants are asked whether the orientation of a visual target bar is the same or different between two arrays separated by a screen break with a fixation cross. We manipulated three aspects of the task: perceptual load, low (displaying 4 objects on the screen; see Figure 1.B) or high (displaying 6 objects on the screen; see Figure 1.B), the presence of audiovisual CC (modality conditions) during different phases of the task (see Figure 2.A-E) and whether the participants were instructed to attend to both visual and auditory stimuli or only the visual stimuli. All participants performed the task in 'High load' and 'Low load' conditions and were exposed to all the modality conditions, while we split participants into two groups and randomly assigned to the 'Attend Visual' or 'Attend Both' conditions.

Figure 1

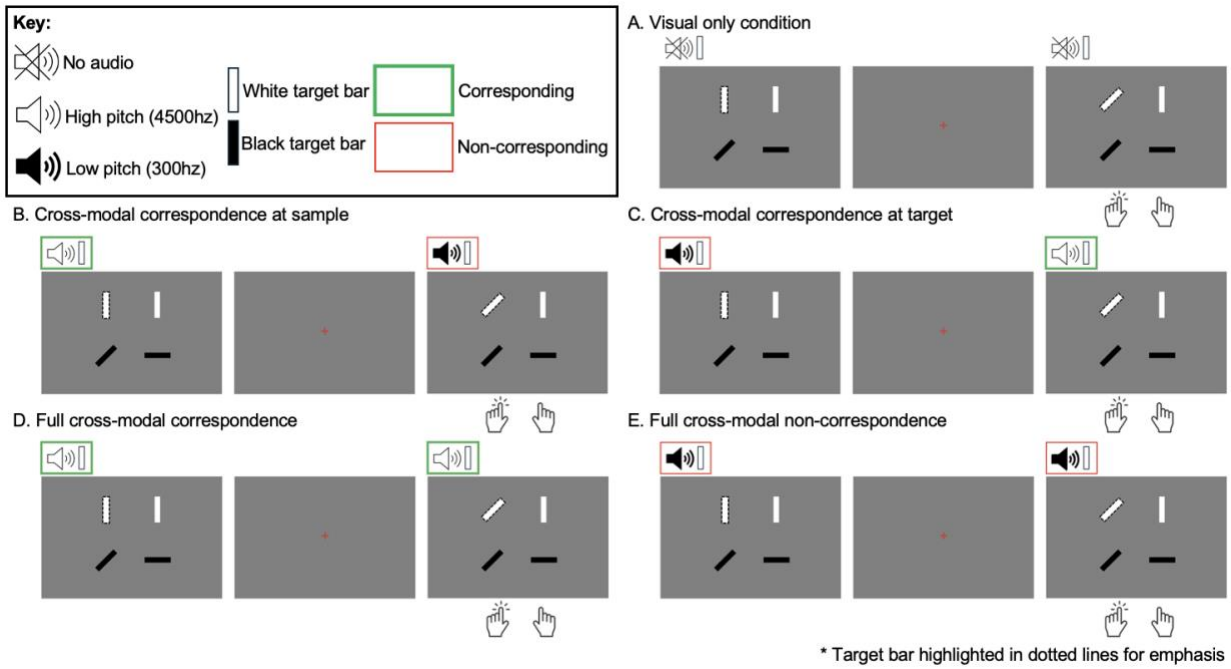
The Low load and High load versions of the orientation change detection task.



Note. Sample: sample array (memory encoding phase); Display until response: target array (memory recall phase).

Figure 2

Modality conditions



The visual stimuli consisted of black and white bars (size 0.10-0.01 height units) on a uniform grey (50%) background. The 'Low load' condition of the task consisted of the presentation of 4 bars (positioned respectively at bar 1) –0.10, 0.10; bar 2) 0.10, 0.10; bar 3) -0.10, -0.10; bar 4) 0.10, -0.10 height units; see Figure 1.A), while the 'High load' condition consisted of the presentation of 6 bars (positioned at bar 1) –0.10, 0.10; bar 2) 0, 0.15; bar 3) 0.10, 0.10; bar 4) -0.10, -0.10; bar 5) 0, -0.15; bar 6) 0.10, -0.10 height units see Figure 1.B). Each bar orientation could vary between 45, 90, 135 and 360 degrees of visual angle. The auditory stimuli consist of either low (300hz) or high (4500hz) pitched sounds presented for 150 ms at the onset of the sample array (encoding phase) and the target array (recall phase).

Audiovisual CC was created by utilizing lightness (black vs white) / pitch (low vs high) correspondence between visual and auditory stimuli. Previous research (Zeljko et

al., 2019) has shown that when audiovisual stimuli varied between two values along each sensory dimension (white or black in the visual dimension and low-pitched or high-pitched in the auditory dimension), participants performed better and faster when the stimuli were paired as white and high-pitched or black and low-pitched, compared with the alternative pairing. Following this evidence, in our experiment, we manipulated the visual stimuli based on lightness (white or black) and the auditory stimuli based on pitch (high or low). The audiovisual correspondence was created on the visual target stimulus (e.g., on the bar that would be presented at a different position during the sample and target array when those were different). Following Zeljko et al. (2019), audiovisual CC was present when either a white target bar was presented with a high-pitched sound (4500hz), or a black target bar was presented with a low-pitched sound (300hz). On the other hand, the absence of audiovisual CC was obtained by presenting together a high-pitched sound (4500hz) with a black target bar or a low-pitched sound (300hz) with a white target bar. We presented audiovisual correspondence during the different stages of memory processing, i.e., at the onset of the sample array (encoding phase) and the target array presentation (recall phase).

So, the task has five conditions pertaining to the presence or absence of audiovisual CC during different phases of the task: ‘Visual-only’ (without sound), ‘CC at sample’ (audiovisual CC during the sample array and non-corresponding stimuli during the target array), ‘CC at target’ (audiovisual CC during the target array and non-corresponding stimuli during the sample array), ‘CC at both’ (audiovisual CC during both arrays), ‘CC at neither’ (non-corresponding audiovisual pairing during both arrays), (see Figure 2). There were two conditions pertaining to the auditory stimuli's attendance

during the orientation change detection task. In the ‘Attend visual’ condition, participants were instructed to pay attention only to the visual stimuli during the task (instructions: “Please only pay attention to what you see on the screen and try to ignore the sound”). In the ‘Attend both’ condition, participants were instructed to pay attention to both the visual and auditory stimuli (instructions: “Please pay attention to both the sound and what you see on the screen”).

Participants were required to use a computer for the study, be in a well-lit place, away from distraction, sit at a 50-80 cm distance from the computer screen, and use a set of headphones or earphones. Before the beginning of the study, participants were instructed to set the volume so that the tone was clearly audible but comfortable. To ensure that the volume was on and to avoid BOT performing the task, participants were asked to listen and identify a brief piece of music (“Happy Birthday”).

Design and Procedure

The study followed a mixed factorial design. The presence or absence of audiovisual cross-modal correspondence during different phases of the task was set as the within-subject factor. The attendance of visual stimuli or both visual and auditory stimuli was set as between-subject factor. The 'Low load’ and ‘High load’ conditions of the task were analyzed independently from each other, to determine the role perceptual load plays in the effect of CC’s on VWM.

In the orientation change detection task, participants were asked to determine whether the orientation of the bars on the target array were the same or different compared to the sample array. Participants answered ‘Same’ by pressing the ‘L’ key, and ‘Different’ by pressing the ‘A’ key on the computer keyboard. The orientation of one

bar on the target array changed 50% of the time. Before each trial, a fixation cross appeared for 600ms, followed by the sample array, paired with a synchronous corresponding or non-corresponding pitch in the cross-modal conditions (150ms), followed by a delay interval with a fixation cross (900ms), followed by the target array (until response or up to 5 seconds) paired with a synchronous corresponding or non-corresponding pitch in the cross-modal conditions (150ms). Each trial had an interstimulus interval of 1800ms (see Figure 1).

Before beginning the experiment, a training round for each block, consisting of 10 'Visual-only' trials without sound and 10 audiovisual trials, was administered in this order. Participants were provided feedback during the training, but no error feedback was provided during the experimental trials. The experiment took approximately 15 minutes to complete. Participants were semi-randomly (controlling for sex) assigned to one of two groups. The first group, 'Attend visual' (n = 21 'Low load' condition; n = 21 'High load' condition) were instructed to perform the change detection task focusing only on the visual stimuli while ignoring the auditory stimuli; the second group, 'Attend both' (n = 22 'Low load' condition; n = 20 'High load' condition) were instructed to pay attention to both the visual stimuli and the sound. Each participant, regardless of their group, was subjected to the same load and audiovisual correspondence conditions. The participants performed 20 trials of the 'Visual-only' condition first and then 80 trials of the four modality conditions in random order. The 'Low load' condition of the task was presented first.

Analysis

Our study aimed at assessing how modality conditions (with the presence or absence of CC) at different task phases ('Visual-only', 'CC at sample', 'CC at target', 'CC at both', 'CC at neither') and attended sensory modality ('Attend Visual', 'Attend both') impacted participants' performance during a change detection task under different perceptual load conditions (low and high). We evaluated performance by looking at accuracy and RT as dependent variables.

For accuracy and RT in each load condition, a 5 × 2 mixed analysis of variance (ANOVA) was conducted. The modality conditions ('Visual-only', 'CC at sample', 'CC at target', 'CC at both', 'CC at neither') were set as within-subjects factor, and the attended modality ('Attend Visual', 'Attend both') was set as between-subjects factor. So, a total of four 5 × 2 ANOVAs were performed, two for the low perceptual load condition and two for the high perceptual load condition. When RT was set as the dependent variable, only RTs during accurate trials were included in the analysis.

Results

VWM under the Low load condition

Accuracy

A 2 × 5 ANOVA was conducted with the modality conditions set as within-subject factor ('Visual-only', 'CC at sample', 'CC at target', 'CC at both', 'CC at neither') and attended modality ('Attend Visual', 'Attend both') set as between-subject factor. Detection of target orientation accuracy was the dependent variable. There was a significant main effect of the modality conditions on accuracy $F_{(4, 164)} = 6.604$, $p < .001$,

ηp^2 = .139. The main effect was explored by using post-hoc repeated measures t-tests, to which Bonferroni correction was applied. Only results with $p < .005$ ($p < .05/10 = .005$; corrected $p < .05$) were considered significant. The accuracy during the 'Visual-only' condition resulted lower than the accuracy during every audiovisual condition: 'Visual-only' and 'CC at sample' $t_{(42)} = -3.807$, $p < .001$, 'Visual-only' and 'CC at target' $t_{(42)} = -3.229$, $p = .002$, 'Visual-only' and 'CC at both' $t_{(42)} = -3.642$, $p < .001$, 'Visual-only' and 'CC at neither' $t_{(42)} = 3.120$, $p = .003$. These findings indicate that the presence of sound, and not the presence of CC, improved participants' performance in the 'Low load' condition (see Table 1 and Figure 3-A.2).

There was no significant main effect of attended modality $F_{(1, 41)} = .003$, $p = .954$, $\eta p^2 = .000$, nor significant interaction between attended modality and the modality conditions $F_{(4,164)} = 2.193$ $p = .072$, $\eta p^2 = .051$.

Table 1

Post-hoc paired sample t-tests for modality conditions: Visual-only, CC at sample, CC at target, CC at both, and CC at neither during the Low load condition.

Condition	t	df	p
Visual-only * CC at sample	-3.807	42	**< .001**
Visual-only * CC at target	-3.229	42	**.002**
Visual-only * CC at both	-3.642	42	**< .001**
Visual-only * CC at neither	-3.120	42	**.003**
CC at sample * CC at target	.527	42	.601
CC at sample * CC at both	-.122	42	.903
CC at sample * CC at neither	.301	42	.765
CC at target * CC at both	-.835	42	.409
CC at target * CC at neither	-.324	42	.747
CC at both * CC at neither	.423	42	.674

Note. N = 43. Bonferroni correction for multiple comparison was applied: results are considered significant when p < .005 (p < .05/10 = .005; corrected p < .05). Significant effects are shown in bold.

Table 2

Means and standard deviations of accuracy during different modality conditions for both between-subject groups during the Low load condition.

	Accuracy					
	Visual-only	**CC at sample**	**CC at target**	**CC at both**	**CC at neither**	**Total**
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Attend visual	.735 (.183)	.868 (.157)	.879 (.128)	.907 (.121)	.886 (.126)	.855 (.143)
Attend both	.813 (.207)	.887 (.153)	.850 (.146)	.855 (.151)	.858 (.120)	.853 (.155)
Total	.775 (.197)	.878 (.153)	.864 (.137)	.881 (.138)	.871 (.122)	.872 (.149)

Note. N = 43. Attend visual group n = 21. Attend both auditory and visual group n = 22.

Reaction time

A 2 × 5 mixed factorial ANOVA was performed, with the modality conditions set as within-subject factor (‘Visual-only’, ‘CC at sample’, ‘CC at target’, ‘CC at both’, ‘CC at neither’) and attended modality set as between-subject factor ('Attend visual’, ‘Attend both’). RT for the accurate trials was set as the dependent variable. The sphericity was violated; therefore, the Greenhouse Geiser correction was applied when reporting the findings. There was a significant main effect of the modality conditions on RT $F_{(3.044,\ 124.784)} = 2.962$, $p = .034$, $\eta_p^2 = .067$. The main effect was explored by using post-hoc repeated measures t-tests, to which Bonferroni correction was applied. Only results with $p < .005$ ($p < .05/10 = .005$; corrected $p < .05$) were considered significant. No significant differences were found between the modality conditions (see Table 3 and Figure 3-B.2). This suggests the ANOVA's significant result is likely due to the overall pattern of variation rather than specific group differences. There was no significant main effect of attended modality $F_{(1,\ 41)} = 2.800$, $p = .102$, $\eta_p^2 = .064$, nor significant interaction between attended modality and the modality conditions $F_{(3.044,\ 124.784)} = 1.726$, $p = .164$, $\eta_p^2 = .040$.

Table 3

Post-hoc paired sample t-tests for modality conditions: Visual-only, CC at sample, CC at target, CC at both, and CC at neither during the Low load condition.

CC condition	t	df	p
Visual-only * CC at sample	2.234	42	.031
Visual-only * CC at target	1.663	42	.104
Visual-only * CC at both	1.461	42	.151
Visual-only * CC at neither	2.774	42	.008
CC at sample * CC at target	-.956	42	.344
CC at sample * CC at both	-.1020	42	.313
CC at sample * CC at neither	.320	42	.751
CC at target * CC at both	-.204	42	.839
CC at target * CC at neither	1.512	42	.138
CC at both * CC at neither	1.322	42	.193

Note. N = 43. Bonferroni correction for multiple comparison was applied: results are considered significant when p < .005 (p < .05/10 = .005; corrected p< .05).

Table 4

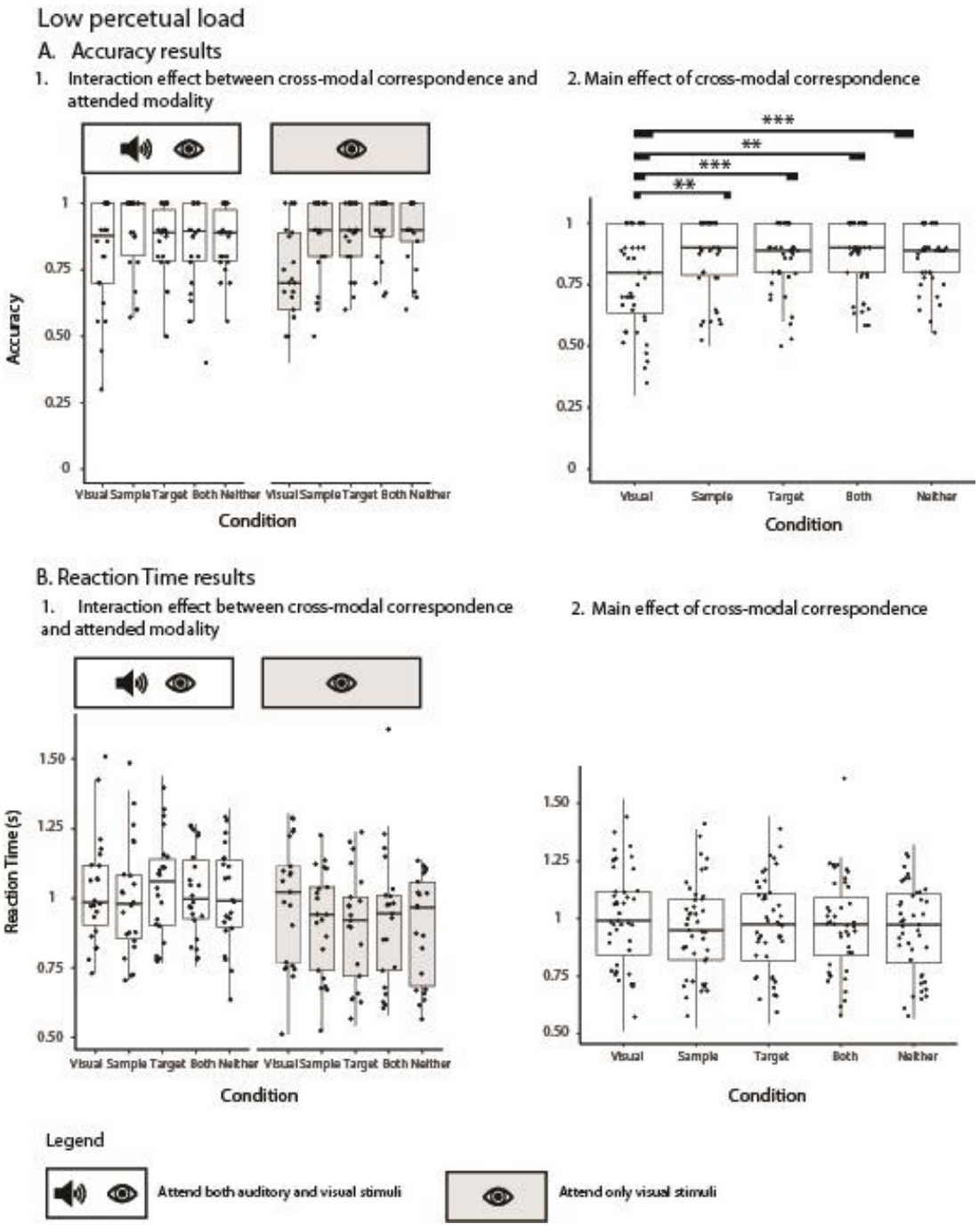
Means and standard deviations of RT during different modality conditions for both between-subject groups during the Low load condition (4-object).

	RT					
	Visual-only	**CC at sample**	**CC at target**	**CC at both**	**CC at neither**	**Total**
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Attend visual	.982 (.224)	.913 (.190)	.899 (.204)	.923 (.247)	.891 (.195)	.922 (.212)
Attend both	1.029 (.195)	.993 (.207)	1.035 (.186)	1.019 (.161)	1.001 (.181)	1.015 (.186)
Total	1.001 (.208)	.954 (.201)	.969 (.204)	.972 (.211)	.950 (.195)	.970 (.204)

Note. N = 43. Attend visual group n = 21. Attend both auditory and visual group n = 22.

Figure 3

Box plots show the accuracy and RT during the low perceptual load condition.



Note. The graphs show the interaction effects between modality conditions and attended modality for accuracy (A1) and RT (B1) and the main effects of modality conditions for accuracy (A2) and RT (B2). The main effect of modality condition is significant both for accuracy ($p < .001$) and RT ($p = .034$); however, post-hoc t-tests to which Bonferroni correction was applied did not show any significant comparison between the conditions when RT was set as dependent variable.

Visual: Visual-only condition; Sample: CC at Sample; Target: CC at Target; Both: CC at Both Target and Sample; Neither: absence of CC at both Target and Sample. *** $p<0.001$; ** $p<.005$. Bonferroni correction has been applied to the comparisons between CC conditions. Only comparisons having a $p<.005$ have been considered significant.

VWM under the High load condition

Accuracy

Consistent with the analysis for the 'Low load' condition, a 2 × 5 ANOVA was conducted with the modality conditions set as within-subject factor ('Visual-only', 'CC at sample', 'CC at target', 'CC at both', 'CC at neither') and attended modality ('Attend Visual', 'Attend both') set as between-subject factor. Detection of target orientation accuracy was the dependent variable. The sphericity was violated; therefore, the Greenhouse Geiser correction was applied when reporting the findings. There was a significant main effect of the modality conditions on accuracy $F_{(3.305,\ 128.907)} = 8.560$, $p < .001$, $\eta p^2 = .180$. The main effect was explored by using post-hoc repeated measures t-tests, to which Bonferroni correction was applied. Only results with $p < .005$ ($p < .05/10 = .005$; corrected $p < .05$) were considered significant. Participants' accuracy was lower during the 'Visual-only' condition compared to the conditions with audiovisual correspondence: 'Visual-only' and 'CC at sample' $t_{(40)} = -3.443$, $p < .001$, 'Visual-only' and 'CC at target' $t_{(40)} = -3.909$, $p < .001$, 'Visual-only' and 'CC at both' $t_{(40)} = -4.468$,

p < .001. We also found that participants were more accurate during the 'CC at both' condition compared to the 'CC at neither' condition $t_{(40)} = 3.117$, p = .003. These findings indicate that the presence of CC, especially when presented both at memory encoding (during the presentation of the sample array) and at memory recall (during the presentation of the target array) improved participants' performance in the 'High load' condition (see Table 5 and Figure 4-A.2). There was no significant main effect of attended modality $F_{(1, 39)} = .486$, p = .490, $\eta p^2 = .012$, nor significant interaction between attended modality and the modality conditions $F_{(3.305, 128.907)} = .724$, p = .552, $\eta p^2 = .018$.

Table 5

Post-hoc paired sample t-tests for modality conditions: Visual-only, CC at sample, CC at target, CC at both, CC at neither during the High load condition.

CC condition	t	df	p
Visual-only * CC at sample	-3.443	40	**.001**
Visual-only * CC at target	-3.909	40	**<.001**
Visual-only * CC at both	-4.468	40	**<.001**
Visual-only * CC at neither	-2.848	40	.007
CC at sample * CC at target	-1.507	40	.140
CC at sample * CC at both	-2.447	40	.019
CC at sample * CC at neither	-.383	40	.704
CC at target * CC at both	-1.038	40	.305
CC at target * CC at neither	1.503	40	.141
CC at both * CC at neither	3.117	40	**.003**

Note. N = 41. Bonferroni correction for multiple comparison was applied: results are considered significant when p < .005 (p < .05/10 = .005; corrected p< .05). Significant effects are shown in bold.

Table 6

Means and standard deviations of accuracy during different modality conditions for both between-subject groups during the High load condition.

	Accuracy					
	Visual-only	**CC at sample**	**CC at target**	**CC at both**	**CC at neither**	**Total**
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Attend visual	.730 (.209)	.797 (.155)	.851 (.160)	.877 (.112)	.832 (.154)	.817 (.158)
Attend both	.699 (.213)	.812 (.152)	.826 (.164)	.858 (.133)	.754 (.208)	.789 (.174)
Total	.715 (.209)	.804 (.151)	.839 (.160)	.868 (.122)	.794 (.184)	.804 (.165)

Note. N = 41. Attend visual group n = 21. Attend both auditory and visual group n = 20.

Reaction time

A 2 × 5 ANOVA was conducted with the modality conditions set as within-subject factor ('Visual-only', 'CC at sample', 'CC at target', 'CC at both', 'CC at neither') and attended modality ('Attend visual', 'Attend both') set as between-subject factor. RT for the accurate trials was set as the dependent variable. The sphericity was violated; therefore, the Greenhouse Geiser correction was applied when reporting the findings. There was a significant main effect of the modality conditions on RT $F_{(2.856, 111.367)} = 4.598$, $p = .005$, $\eta_p^2 = .105$. The main effect was explored by using post-hoc repeated measures t-tests, to which Bonferroni correction was applied. Only results with $p < .005$ ($p < .05/10 = .005$; corrected $p < .05$) were considered significant. Significant differences were found between the 'Visual-only' condition and the 'CC at

both' condition: $t_{(40)} = 3.811$, $p < .001$. These findings indicate that CC presented both at memory encoding (at the presentation of the sample array) and at memory recall (at the presentation of the target array) improved participants' performance compared to the 'Visual-only' condition. However, no significant differences were found between cross-modally corresponding and non-corresponding conditions (see Table 7 and Figure 4-B.2). There was no main effect of attended modality $F_{(1, 39)} = 1.645$, $p = .207$, $\eta_p^2 = .040$, nor significant interaction between attended modality and the modality conditions $F_{(2.856, 111.367)} = 1.456$, $p = .232$, $\eta_p^2 = .036$.

Table 7

Post-hoc paired sample t-tests for modality conditions: Visual-only, CC at sample, CC at target, CC at both, CC at neither during the High load condition.

CC condition	t	df	p
Visual-only * CC at sample	1.876	40	.068
Visual-only * CC at target	2.921	40	.006
Visual-only * CC at both	3.811	40	**<.001**
Visual-only * CC at neither	1.807	40	.078
CC at sample * CC at target	.985	40	.331
CC at sample * CC at both	1.830	40	.075
CC at sample * CC at neither	-.153	40	.879
CC at target * CC at both	.898	40	.375
CC at target * CC at neither	-1.205	40	.235
CC at both * CC at neither	-1.741	40	.089

Note. N = 41. Bonferroni correction for multiple comparison was applied: results are considered significant when $p < .005$ ($p < .05/10 = .005$; corrected $p < .05$). Significant effects are shown in bold.

Table 8

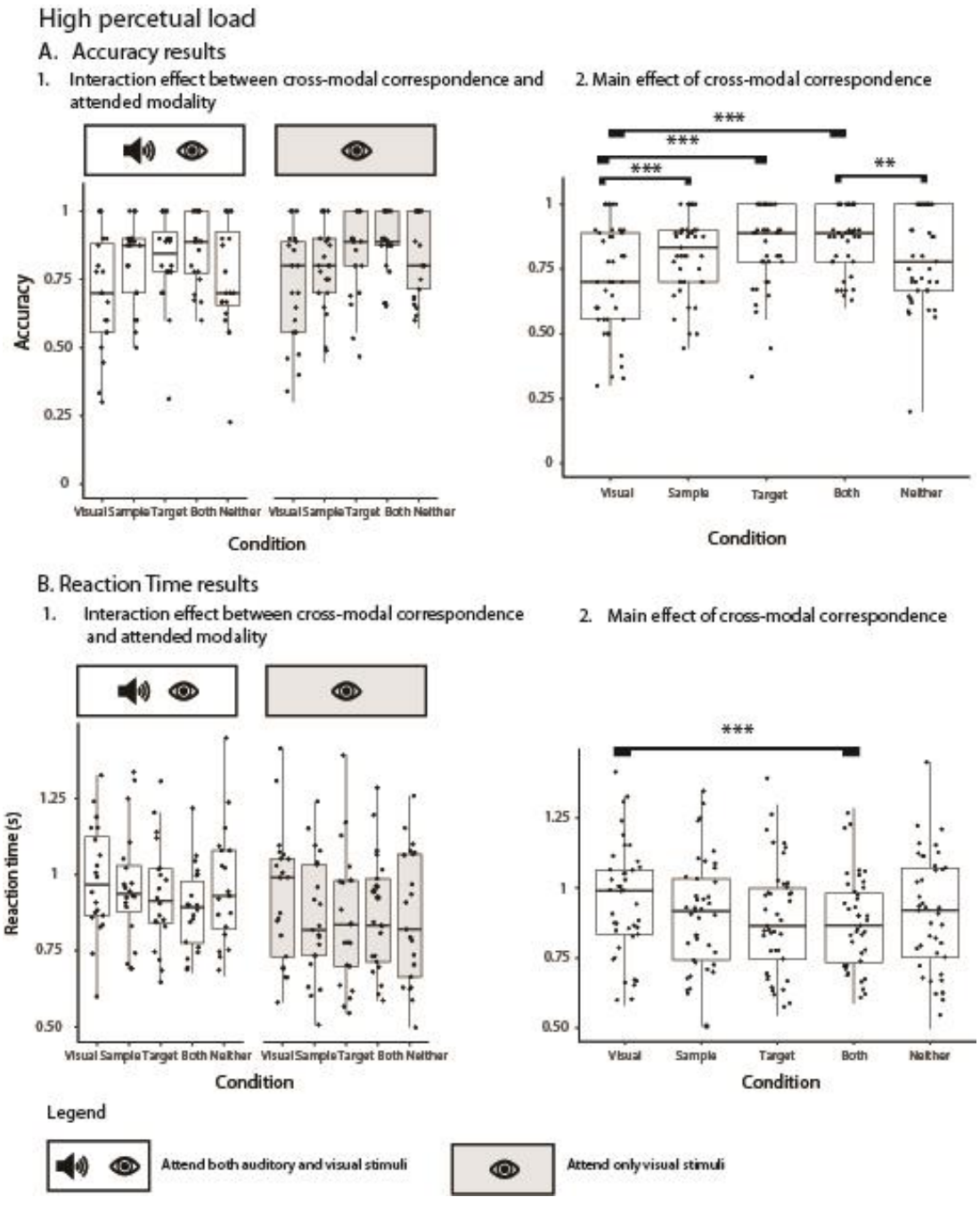
Means and standard deviations of RT during different modality conditions for both between-subject groups during the High load condition.

	RT					
	Visual-only	**CC at sample**	**CC at target**	**CC at both**	**CC at neither**	**Total**
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Attend visual	.927 (.218)	.856 (.198)	.852 (.221)	.865 (.194)	.861 (.219)	.872 (.210)
Attend both	.980 (.181)	.958 (.187)	.930 (.173)	.889 (.139)	.958 (.190)	.943 (.174)
Total	.953 (.200)	.906 (.197)	.890 (.200)	.877 (.168)	.908 (.209)	.923 (.195)

Note. N = 41. Attend visual group n = 21. Attend both auditory and visual group n = 20.

Figure 4

Box plots show the accuracy and RT during the high perceptual load condition.



Note. The graphs show the interaction effects between modality conditions and attended modality for accuracy (A1) and RT (B3) and the main effects of modality conditions for accuracy (A2) and RT (B2).

Visual: Visual-only condition; Sample: CC at sample; Target: CC at target; Both: CC at both target and sample; Neither: absence of CC at both target and sample. *** p<.001; ** p<.005. Bonferroni correction has been applied to the comparisons between CC conditions. Only comparisons having a p<.005 have been considered significant.

Discussion

The present study investigated the effect of lightness/pitch correspondence on VWM performance testing the impact of perceptual load and instructed attended modality. We designed a black-and-white orientation change detection task where visual stimuli were paired with high- and low-pitched sounds. To manipulate perceptual load, the experiment consisted of a 4-object condition ('Low load') and a 6-object condition ('High load'). We instructed a group of participants only to pay attention to the visual stimuli and ignore the sound, while the other group was asked to pay attention to both visual and auditory stimuli. Our findings provide several critical insights into the dynamics of multisensory integration and its impact on WM processes, revealing that the cognitive system benefits from lightness/pitch correspondence under increased perceptual load, and this process is not influenced by conscious attentional focus to simultaneous auditory input.

Our results suggest that, in the lightness/pitch domain, CC does not automatically enhance performance on the orientation change detection task; instead, the effect is modulated by perceptual load. Under high perceptual load, when compared to the 'Visual-only' condition, participants exhibited significantly enhanced accuracy in the presence of corresponding audiovisual stimuli. This effect was consistently observed when CC was displayed during memory encoding, memory recall, and when it was displayed during both stages. Furthermore, when CC was presented during both

memory encoding and memory recall within a trial, it also significantly improved accuracy compared to the non-corresponding audiovisual condition. This result aligns with prior research by Brunetti et al. (2017), who reported a marginally significant effect suggesting that CC (e.g., elevation/pitch, shape/pitch, audiovisual numerosity) can improve WM accuracy in the n-back task. Specifically, this finding was observed when CC was present during both the target array and sample array, compared to the fully non-corresponding condition. Similarly, Constant and Liesefeld (2020) demonstrated that stimulus saliency has a strong effect on VWM performance. Our findings (see Figure 4) revealed that this effect is driven more strongly by the memory recall phase, even in trials where corresponding stimuli were presented during both encoding and recall stages. Specifically, CC presented during memory recall produced a similar pattern to CC presented during both memory encoding and recall.

Under high perceptual load, presenting CC during both memory processing stages also led to significantly faster RT compared to the conditions where only visual stimuli were presented. This finding is consistent with Brunetti et al. (2017), who demonstrated that various cross-modal correspondences (e.g., elevation/pitch, shape/pitch, audiovisual numerosity) can enhance RT. We also observed a similarity in the RT data pattern between CC at memory recall and CC at both stages (see Figure 4). Our results show that, under high perceptual load, corresponding audiovisual stimuli presented during both encoding and recall stages led to the highest accuracy and fastest RT. Having CC present during either encoding or recall alone also enhanced performance, albeit to a lesser extent than when CC was present during both encoding and recall stages. These findings point towards cross-modal sensory integration playing

a role throughout the entire VWM process in a distributed nature (Christophel et al., 2017; McEwan et al., 2024).

Conversely, under ‘Low load’ conditions, the presence of auditory stimuli, whether corresponding or not, significantly improved accuracy compared to the ‘Visual-only’ condition, but no significant differences were observed between corresponding and non-corresponding audiovisual conditions. These findings indicate that under low perceptual load, the cognitive system benefits from the general arousal and alertness effects induced by auditory stimuli (Han et al., 2013; Han et al., 2021) without a specific advantage for CC. Our results suggest that the brain utilizes corresponding cross-modal information to enhance VWM performance when attentional resources are strained. This suggests that strong associations between audiovisual stimuli play a crucial role in withstanding the effects of increased perceptual load on VWM processing.

The evidence that load modulates the effect of CC at different stages of WM processing aligns with Lavie's load theory (Lavie, 1995; Lavie et al., 2004), which proposes that high perceptual load restricts the processing of contextually irrelevant information. Our load specific results also corroborate with Li et al. (2022), who demonstrated that perceptual load plays a critical role in multisensory integration; higher loads can enhance the processing of congruent audiovisual stimuli while suppressing the processing of incongruent ones. An EEG study by Simon et al. (2016) demonstrated similar results regarding WM load. The authors found that increased WM load is linked to stronger and more focused attention on the primary visual task, leading to reduced processing of cross-modal auditory stimuli that are irrelevant to the task. The attentional effect of CC is well established by prior research (Brunetti et al., 2017; Chiou & Rich,

2012; Klapetek et al., 2012). Similarly to Brunetti et al. (2017), our results show that the attentional effect evoked by CC affects not only the WM encoding stage but also the recall phase.

Our findings contribute to the ongoing debate about the conditions under which sound facilitates VWM performance and when it becomes a distractor. Our results suggest that the interplay between audiovisual CC and perceptual load can be a key factor in determining whether auditory information serves as a beneficial cue (e.g., through contextual or attentional cueing) or a distracting element (e.g., through irrelevance). It is a widely accepted view that general WM capacity is 4±1 items under typical conditions (Cowan, 2010), and VWM capacity is estimated to be 3-4 items (Dai et al., 2019; Luck & Vogel, 2013). Research (Lavie et al., 2004; De Fockert et al., 2001) indicates that high perceptual load (the complexity or quantity of stimuli) decreases susceptibility to distractions; however, high WM load (the amount of information to be held and manipulated) increases it. Even so, Simon et al. (2016) found that during audiovisual cross-modal tasks high WM load may reduce the processing of task-irrelevant auditory stimuli through focused attention. Furthermore, according to Liesefeld et al. (2020) VWM capacity is aided by its ability to filter out irrelevant information and selectively encode relevant information. These findings suggest that WM's filtering ability may be influenced by contextual associations between stimuli, particularly when perceptual load is high.

Chunking (i.e., transforming smaller pieces of information into larger familiar units based on contextual association) is often believed to help overcome the limited capacity of WM. Thalmann et al. (2019) suggested that chunks reduce WM load through memory

retrieval of compact chunk representations, thereby replacing the representations of individual elements of the chunk. This account is supported by Son et al. (2020), who found that VWM clusters sensory items into larger representational units based on similarity. CC can be considered a form of chunking or clustering. Brunel et al. (2015) demonstrated that the association between sensory information from different modalities modulates cross-modal integration during perceptual learning, leading to newly learned units. Clustering information based on CC into fewer units, when attentional resources are stretched, could explain why high perceptual load does not necessarily increase cognitive load, as suggested by Lavie (1995; 2005). Consequently, our findings support this notion. Merging corresponding stimuli into cohesive units, combined with the attentional effect of CC (Brunetti et al., 2017; Chiou & Rich, 2012; Klapetek et al., 2012), makes it more likely that visual items associated with the auditory stimuli are prioritized for processing when attentional resources are stretched.

Orbán et al. (2008) suggested that people use Bayesian learning to form efficient visual chunks from complex patterns based on statistical regularities in the environment. Statistical learning is the process through which the brain automatically and unconsciously learns the relationships between stimuli based on associations (Barakat et al., 2012). In addition, Bayesian learning also uses prior knowledge to predict sensory information and updates beliefs based on prediction errors (Shams & Beierholm, 2022). Prior studies have shown a link between statistical and Bayesian learning and VWM. Brady et al. (2009) found that participants' memory for the colors of concentric circle pairs improved when the colors were correlated (consistent color pairing patterns across trials) but dropped to control levels when correlations were

removed. Bates et al. (2019) explored how the brain leverages statistical regularities in visual environments to form efficient VWM representations, finding that participants implicitly learn and use these regularities to enhance VWM performance. Moreover, a study by Umemoto et al. (2010) demonstrates that statistical learning may help to optimize the allocation of limited resources in WM by biasing encoding towards behaviorally relevant items.

According to Spence (2011), certain cross-modal correspondences can be understood in terms of Bayesian learning. This framework suggests that people may integrate stimuli in a statistically optimal way by combining sensory information and prior knowledge (also called priors) and weighting each of them by their relative reliabilities (Ernst, 2007; Parise & Spence, 2009). According to this model, the brain creates connections (or couplings) between stimuli to adapt to different situations and constraints. The more prior knowledge the brain has about the association between two stimuli, the stronger the coupling will be (Ernst, 2007). This means that with stronger coupling, unisensory signals from multimodal sources are more likely to merge into a single multisensory unit (Spence, 2011).

Previous research suggests that Bayesian models can support both parallel and integrated WM representations (Ma et al., 2006; Shams & Beierholm, 2022). This study does not aim to determine if lightness/pitch stimuli are maintained separately in WM by sensory modality or recoded into multisensory representations. However, we hypothesize that when sensory information is processed based on associations, CC at memory recall may facilitate integrating parallel WM representations into a single memory response through attentional cueing. This integration can be achieved by

653 treating recall as a Bayesian inference process, where chunk representations provide
654 priors to interpret sensory input, as suggested by Norris and Kalm (2021).

655 Based on our findings, we propose that, as a dynamic adaptation to manage the
656 increase in perceptual load, the cognitive system may rely more heavily on organizing,
657 prioritizing, encoding, and recalling information based on learned statistical patterns.
658 Clustering crossmodally corresponding stimuli into fewer multisensory units based on
659 highly probabilistic associations could be one such mechanism that aids WM processes
660 when attentional resources are strained. Inference about whether cross-modal stimuli
661 are correlated could also aid in filtering out irrelevant sensory information and prioritizing
662 the selection of relevant information under increased perceptual load. Additionally, our
663 results suggest that immediate priors, such as displaying corresponding stimuli during
664 both memory encoding and recall, can further strengthen these associations and
665 thereby aid VWM performance.

666 Contrary to our hypothesis, the attended modality (whether participants paid
667 attention to both visual and auditory stimuli or only visual stimuli) did not significantly
668 impact accuracy or RT. This suggests that the benefits of lightness/pitch CC on VWM
669 are robust and not heavily dependent on top-down attentional control. Instead, the
670 effects appear to be automatically driven by the characteristics of the cross-modal
671 stimuli and their integration. This finding is consistent with previous research indicating
672 that multisensory binding can occur automatically and does not always require focused
673 attention (Molholm et al., 2007; Zlejko et al., 2019, 2021). This automaticity can also be
674 explained by Hebb's law (Hebb, 1949). The repeated co-activation of neurons
675 responding to visual and auditory stimuli during CC strengthens their connections. Due

to this strengthening, even when attention is only directed to one of these modalities, the neural pathways linking them may operate automatically. Spence (2011) suggested that some CCs involving pitch are likely statistical correspondences representing the internalization of the natural correlations between stimulus attributes present in the environment. Furthermore, any unimodal component of such multisensory pair can sufficiently activate the association representing the other unimodal component. Expanding on these findings our study indicates that lightness/pitch CC could operate in a similar fashion.

The present study provides key critical insights into the impact of multisensory integration on VWM. However, some limitations should be noted, together with future research directions. The present study only included two load conditions and only explored lightness/pitch CC. Future research could investigate whether higher perceptual load increases the distractor effect of non-corresponding pitch on VWM and whether this effect is modulated by the number of stimuli included in the set through a gradient effect (e.g., the advantage is increased gradually as the number of stimuli in the set increases) or whether the effect is modulated in an on/off manner (e.g., the advantage is present for sets with a certain number of stimuli, but it does not increase when the number of stimuli increases after this number). Another opportunity for further studies is to explore whether the findings are generalizable to other types of CC correspondences, (e.g., auditory pitch/visual shape, numerosity, phonetic characteristics of the auditory stimulus/ visual shape, auditory frequency/ visual frequency) and other sensory modalities (e.g., in the tactile domain). In the current study, participants were required only to perform the WM task in the visual domain. The

saliency of the stimulus modality could have been modulated by the relevance of the type of stimuli for performing task. Future research that requires tasks to be performed either in the visual or in the auditory domain could provide further insights into how the interaction between attention and cross-modal correspondence affects the larger WM system and its domain-specific and domain-general networks (see, e.g., Li et al., 2014). A further avenue of investigation is the neural mechanisms underlying the interplay between CC and perceptual load on WM processing using neuroimaging techniques. Analytical approaches based on machine learning as Multi-Voxel-Pattern-Analysis (Norman et al., 2006) or Representational Similarity Analysis (Kriegeskorte et al., 2008) could help to shed a light on the nature of neural representations of cross-modal correspondences in the sensory and cognitive cortices.

Conclusion

In conclusion, the current study found that, in the lightness/pitch domain, audiovisual CC does not automatically facilitate VWM performance; instead, the salience of the effect is modulated by perceptual load. Lightness/pitch correspondence only improves accuracy and RT under high perceptual load. Contrarily, under low perceptual load, the mere presence of auditory pitch, regardless of visual correspondence, enhances VWM performance, likely by increasing alertness. Moreover, our results demonstrate that lightness/pitch CC is automatic and does not rely on conscious selective attention. Our results also suggest that CC serves a functional role beyond sensory integration. We hypothesize that the cognitive system adapts to high perceptual load by clustering CC stimuli into cohesive units and leveraging principles of statistical learning, primarily through the Bayesian approach.

Our study provides a deeper understanding of how the cognitive system adjusts to perceptual demands by demonstrating the effects of the interplay between lightness/pitch CC and perceptual load on VWM performance. It also lays the foundation for future research and interventions that could optimize cognitive performance across various domains, including education, transport safety, and clinical settings.

References

American Psychological Association (2017). *Ethical Principles of Psychologists and Code of Conduct.* https://www.apa.org/ethics/code

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Badde, S., Navarro, K. T., & Landy, M. S. (2020). Modality-specific attention attenuates visual-tactile integration and recalibration effects by reducing prior expectations of a common source for vision and touch. *Cognition*, *197*, 104170. https://doi.org/10.1016/j.cognition.2019.104170

Barakat, B., Seitz, A., & Shams, L. (2012). There is more to statistical learning than associative learning: Predictable items are enhanced even when not predicted. *Journal of Vision*, *12*(9), 694-694. https://doi.org/10.1167/12.9.694

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, *19*(2), 11-11. https://doi.org/10.1167/19.2.11

Beierholm, U. R., Quartz, S. R., & Shams, L. (2009). Bayesian priors are encoded independently from likelihoods in human multisensory perception. *Journal of Vision, 9*(5), 23. https://doi.org/10.1167/9.5.23

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory

representations. *Journal of Experimental Psychology: General, 138*(4), 487–502. https://doi.org/10.1037/a0016797

British Psychological Society (2021). *BPS Code of Human Research Ethics.* https://www.bps.org.uk/guideline/bps-code-human-research-ethics

Brunel, L., Carvalho, P. F., & Goldstone, R. L. (2015). It does belong together: Cross-modal correspondences influence cross-modal integration during perceptual learning. *Frontiers in Psychology*, *6*, 121086. https://doi.org/10.3389/fpsyg.2015.00358

Brunetti, R., Indraccolo, A., Mastroberardino, S., Spence, C., & Santangelo, V. (2017). The impact of cross-modal correspondences on working memory performance. *Journal of Experimental Psychology: Human Perception and Performance, 43*(4), 819–831. https://doi.org/10.1037/xhp0000348

Burr, D., Banks, M. S., & Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research*, *198*(1), 49–57. https://doi.org/10.1007/s00221-009-1933-z

Butler, J. S., Smith, S. T., Campos, J. L., & Bulthoff, H. H. (2010). Bayesian integration of visual and vestibular signals for heading. *Journal of Vision*, *10*(11), 23–23. https://doi.org/10.1167/10.11.23

Chiou, R., & Rich, A. N. (2012). Cross-Modality Correspondence between Pitch and Spatial Location Modulates Attentional Orienting. *Perception, 41*(3), 339-353. https://doi.org/10.1068/p7161

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J. D. (2017). The distributed nature of working memory. *Trends in Cognitive Sciences*, *21*(2), 111-124. https://doi.org/10.1016/j.tics.2016.12.007

Constant, M., & Liesefeld, H. R. (2020). The role of saliency for visual working memory in complex visual scenes. *Journal of Vision, 20*(11), 499. https://doi.org/10.1167/jov.20.11.499

Cowan, N. (2010). The Magical Mystery Four. *Current Directions in Psychological Science*, *19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., Kawahara, S., Koutalidis, S., Krifka, M., Lippus, P., Lupyan, G., Oh, G. E., Paul, J., Petrone, C., Ridouane, R., Reiter, S., Schümchen, N., Szalontai, Á., Ünal-Logacev, Ö., … Winter, B. (2022). The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1841). https://doi.org/10.1098/rstb.2020.0390

Dai, M., Li, Y., Gan, S., & Du, F. (2019). The reliability of estimating visual working memory capacity. *Scientific Reports*, *9*(1), 1155. https://doi.org/10.1038/s41598-019-39044-1

De Fockert, J.W., Rees, G., Frith, C.D., Lavie, N., 2001. The Role of Working Memory in Visual Selective Attention. *Science, 291*, 1803–1806. https://doi.org/10.1126/science.1056496

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–169. https://doi.org/10.1016/j.tics.2004.02.002

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. https://doi.org/10.3758/bf03193146

Fetsch, C. R., Turner, A. H., Deangelis, G. C., & Angelaki, D. E. (2009). Dynamic Reweighting of Visual and Vestibular Cues during Self-Motion Perception. *The Journal of Neuroscience*, *29*(49), 15601–15612. https://doi.org/10.1523/jneurosci.2574-09.2009

Han, L., Liu, Y., Zhang, D., Jin, Y., & Luo, Y. (2013). Low-Arousal Speech Noise Improves Performance in N-Back Task: An ERP Study. *PLOS ONE*, *8*(10), e76261. https://doi.org/10.1371/journal.pone.0076261

Han, S., Zhu, R., & Ku, Y. (2021). Background white noise and speech facilitate visual working memory. *European Journal of Neuroscience*, *54*(7), 6487-6496. https://doi.org/10.1111/ejn.15455

Hebb, D. O. (1949). *The Organization of Behavior*. https://doi.org/10.4324/9781410612403

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829–6833. https://doi.org/10.1073/pnas.0801268105

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the N-back task: A question of construct validity.

Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(3), 615-622. https://doi.org/10.1037/0278-7393.33.3.615

Klapetek, A., Ngo, M.K., Spence, C., 2012. Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search? *Attention, Perception, & Psychophysics 74*, 1154–1167. https://doi.org/10.3758/s13414-012-0317-9

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249. https://doi.org/10.3389/neuro.06.004.2008

Lamichhane, B., Westbrook, A., Cole, M. W., & Braver, T. S. (2020). Exploring brain-behavior relationships in the N-back task. *Neuroimage*, *212*, 116683. https://doi.org/10.1016/j.neuroimage.2020.116683

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance, 21*(3), 451–468. https://doi.org/10.1037/0096-1523.21.3.451

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences*, *9*(2), 75-82. https://doi.org/10.1016/j.tics.2004.12.004

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load Theory of Selective Attention and Cognitive Control. *Journal of Experimental Psychology: General, 133*(3), 339–354. https://doi.org/10.1037/0096-3445.133.3.339

Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics 56*, 183–197. https://doi.org/10.3758/bf03213897

Liesefeld, H. R., Liesefeld, A. M., Sauseng, P., Jacob, S. N., & Müller, H. J. (2020). How visual working memory handles distraction: cognitive mechanisms and electrophysiological correlates. *Visual Cognition*, *28*(5-8), 372–387. https://doi.org/10.1080/13506285.2020.1773594

Li, D., Christ, S. E., & Cowan, N. (2014). Domain-general and domain-specific functional networks in working memory. *Neuroimage*, *102*, 646–656. https://doi.org/10.1016/j.neuroimage.2014.08.028

Li, Q., Yu, Y., Liu, Y., Xu, Z., Fan, L., Takahashi, S., Yang, J., Ejima, Y., Wu, Q., & Wu, J. (2022). Whether attentional loads influence audiovisual integration depends on semantic associations. *Attention, Perception, & Psychophysics*, *84*(7), 2205–2218. https://doi.org/10.3758/s13414-022-02461-y

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

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438. https://doi.org/10.1038/nn1790

Makovac, E., Kwok, S. C., & Gerbino, W. (2014). Attentional cueing by cross-modal congruency produces both facilitation and inhibition on short-term visual recognition. *Acta Psychologica*, *152*, 75-83. https://doi.org/10.1016/j.actpsy.2014.07.008

878 Marks, L. E. (1987). On cross-modal similarity: Auditory–visual interactions in speeded
879 discrimination. *Journal of Experimental Psychology: Human Perception and*
880 *Performance*, *13*(3), 384. https://doi.org/10.1037/0096-1523.13.3.384
881 McEwan, J., Kritikos, A., & Zeljko, M. (2024). Involvement of the superior colliculi in
882 crossmodal correspondences. *Attention, Perception, & Psychophysics*, *86*(3),
883 931–941. https://doi.org/10.3758/s13414-024-02866-x
884 Molholm, S., Martinez, A., Shpaner, M., & Foxe, J. J. (2007). Object-based attention is
885 multisensory: Co-activation of an object's representations in ignored sensory
886 modalities. *European Journal of Neuroscience*, *26*(2), 499-509.
887 https://doi.org/10.1111/j.1460-9568.2007.05668.x
888 Murray, M. M., Brunet, D., & Michel, C. M. (2008). Topographic ERP Analyses: A Step-
889 by-Step Tutorial Review. *Brain Topography*, *20*(4), 249–264.
890 https://doi.org/10.1007/s10548-008-0054-5
891 Noppeney, U. (2021). Perceptual Inference, Learning, and Attention in a Multisensory
892 World. *Annual Review of Neuroscience*, *44*(1), 449–473.
893 https://doi.org/10.1146/annurev-neuro-100120-085519
894 Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading:
895 multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences, 10*(9),
896 424-430. https://doi.org/10.1016/j.tics.2006.07.005
897 Norris, D., & Kalm, K. (2021). Chunking and data compression in verbal short-term
898 memory. *Cognition*, *208*, 104534. https://doi.org/10.1016/j.cognition.2020.104534

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745–2750. https://doi.org/10.1073/pnas.0708424105

Parise, C. V., & Spence, C. (2009). 'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLOS ONE*, *4*(5), e5664. https://doi.org/10.1371/journal.pone.0005664

Rigo, A., Roebuck, H., & Manini, B. (2024). *Perceptual Load Modulates the Effect of Lightness/Pitch Correspondence on Visual Working Memory Performance*. osf.io/m8t5b

Ross-Sheehy, S., Reynolds, E., & Eschman, B. (2021). Unsupervised online assessment of visual working memory in 4-to 10-year-old children: array size influences capacity estimates and task performance. *Frontiers in Psychology*, *12*, 692228. https://doi.org/10.3389/fpsyg.2021.692228

Rubin, E. (1915). *Synsoplevede Figurer*. Gyldendal.

Simon, S. S., Tusch, E. S., Holcomb, P. J., & Daffner, K. R. (2016). Increasing working memory load reduces processing of cross-modal task-irrelevant stimuli even after controlling for task difficulty and executive capacity. *Frontiers in Human Neuroscience*, *10*, 380. https://doi.org/10.3389/fnhum.2016.00380

Shams, L., & Beierholm, U. (2022). Bayesian causal inference: A unifying neuroscience theory. *Neuroscience & Biobehavioral Reviews*, *137*, 104619. https://doi.org/10.1016/j.neubiorev.2022.104619

Son, G., Oh, B.-I., Kang, M.-S., & Chong, S. C. (2020). Similarity-based clusters are representational units of visual working memory. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 46(1), 46–59. https://doi.org/10.1037/xlm0000722

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971–995. https://doi.org/10.3758/s13414-010-0073-7

Spence, C. (2020). Simple and complex crossmodal correspondences involving audition. *Acoustical Science and Technology*, *41*(1), 6–12. https://doi.org/10.1250/ast.41.6

Spence, C., & Deroy, O. (2013). How automatic are crossmodal correspondences? *Consciousness and Cognition*, *22*(1), 245-260. https://doi.org/10.1016/j.concog.2012.12.006

Spence, C., Senkowski, D., & Röder, B. (2009). Crossmodal processing. *Experimental Brain Research*, *198*(2-3), 107–111. https://doi.org/10.1007/s00221-009-1973-4

Terhune, D. B., Wudarczyk, O. A., Kochuparampil, P., & Kadosh, R. C. (2013). Enhanced dimension-specific visual working memory in grapheme–color synesthesia. *Cognition*, *129*(1), 123-137. https://doi.org/10.1016/j.cognition.2013.06.009

Thalmann, M., Souza, A. S., & Oberauer, K. (2019). How does chunking help working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(1), 37–55. https://doi.org/10.1037/xlm0000578

Umemoto, A., Scolari, M., Vogel, E.K., Awh, E., 2010. Statistical learning induces discrete shifts in the allocation of working memory resources. *Journal of*

Experimental Psychology: Human Perception and Performance 36(6), 1419–1429. https://doi.org/10.1037/a0019324

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the “unity assumption” using audiovisual speech stimuli. *Perception & Psychophysics*, *69*(5), 744–756. https://doi.org/10.3758/bf03193776

Zeljko, M., Grove, P. M., & Kritikos, A. (2021). The lightness/pitch crossmodal correspondence modulates the Rubin face/vase perception. *Multisensory Research, 34*(7), 763–783. https://doi.org/10.1163/22134808-bja10054

Zeljko, M., Kritikos, A., & Grove, P. M. (2019). Lightness/pitch and elevation/pitch crossmodal correspondences are low-level sensory effects. *Attention, Perception, & Psychophysics*, *81*(5), 1609–1623. https://doi.org/10.3758/s13414-019-01668-w

Zhou, Y., Curtis, C. E., Sreenivasan, K. K., & Fougnie, D. (2022). Common Neural Mechanisms Control Attention and Working Memory. *The Journal of Neuroscience*, *42*(37), 7110–7120. https://doi.org/10.1523/jneurosci.0443-22.2022