An Experimental Investigation of Modality Effect: Evidence from Eye-Tracking Data

Inauguraldissertation

der Philosophisch-humanwissenschaftlichen Fakultät

der Universität Bern zur Erlangung der Doktorwürde

Vorgelegt von

Hafidah Binti Umar

von Malaysia

Selbstverlag, Bern, 2020

Original document saved on the web server of the University Library of Bern

This work is licensed under a
Creative Commons Attribution-Non-Commercial-No derivative works 2.5 Switzerland licence.
To see the licence go to http://creativecommons.org/licenses/by-nc-nd/2.5/ch/ or
write to Creative Commons, 171 Second Street, Suite 300, San Francisco, California 94105, USA.

source: https://doi.org/10.24442/boristheses.1774 | downloaded: 20.4.2024

Von der Philosophisch-humanwissenschaftlichen Fakultät der Universität Bern auf Antrag von Prof. Dr. Trix Cacchione (Hauptgutachterin) und Prof. Dr. Fred Mast (Zweitgutachterin) angenommen.

Bern, den 20. Januar 2020 Der Dekan: Prof. Dr. Ernst-Joachim Hossner

Copyright Notice

This document is licensed under the Creative Commons Attribution-Non-Commercial-No derivative works 2.5 Switzerland.
http://creativecommons.org/licenses/by-nc-nd/2.5/ch/

You are free:

to copy, distribute, display, and perform the work

Under the following conditions:

Attribution. You must give the original author credit.

Non-Commercial. You may not use this work for commercial purposes.

No derivative works. You may not alter, transform, or build upon this work.

For any reuse or distribution, you must take clear to others the license terms of this work.

Any of these conditions can be waived if you get permission from the copyright holder.

Nothing in this license impairs or restricts the author's moral rights according to Swiss law.

The detailed license agreement can be found at:
http://creativecommons.org/licenses/by-nc-nd/2.5/ch/legalcode.de

Abstract

Modality effect is a response that occur when there are manipulations of sensory modality. In this thesis, I present a series of studies about the multimodal processing of visual and auditory presentation. The aim of this dissertation is to investigate how would the different stimulations from different source of modalities affect the oculomotor response. I investigate how different stimuli are processed, recognized and retrieved when they are presented across multiple modalities. Specifically, on question of how would the visual and auditory manipulations influence the oculomotor behaviour. In the research area of the multimodal processing, it has been argued that different kinds of sensory manipulations elicit a distinct kind of cognitive and behavioural response. The study of modality effect is particularly interesting topic for investigations since the world is multimodal in nature. Humans and other living beings are constantly exposed to a wide variety of stimuli rather than to isolated single stimulus. All experiments conducted used an eye-tracking approach since eye-tracking data are known as a reliable measure to study implicit cognitive processing. In Experiment 1, I investigate how different modalities and context interplay on the allocation of visual attention during the perceptual processing of congruent and incongruent multimodal stimuli. In Experiment 2, I investigate recognition memory of multimodal stimuli, focusing on the participants' reaction to old versus novel stimuli presented in the visual and auditory modalities. In Experiment 3, I monitored looking patterns, to understand how visual and auditory stimuli are mentally reconstructed during mental imagery. I conclude the dissertation with a discussion of how a different kinds of modality manipulations elicit distinct modality effect as revealed by oculomotor response.

Acknowledgements

I would like to thank:-

My doctoral supervisor Prof. Dr. Trix Cacchione for her support and advice. Thank you for accepting me as a student.

My mentor Prof. Dr. Corinna Martarelli for her guidance, patience and hard work throughout this process.

Prof. Dr. Fred Mast for his expertise and support.

Technical staff at MMZ for their help regarding technical equipment.

All participants in these dissertation studies for their time and effort.

Dr. Nayla Sokhn for her help in statistics with MATLAB and R programming.

Dr. Federica Amici for her encouragement and assistance in academic writing.

Prof. Dato' Dr. Jafri Malin Abdullah from Brain Behaviour Cluster, Department of Neurosciences, Universiti Sains Malaysia for the opportunity to pursue a doctoral study.

Universiti Sains Malaysia and Ministry of Education Malaysia for a fellowship.

My friends for their kindness and understanding.

My parents and all my family members for their kind words and generosity.

My husband for his love and care.

Table of Contents

Abstract 3

Acknowledgements 4

Table of Contents 5

List of Figures 8

List of Tables 9

Chapter 1 10

General Background 10

1.1 Processing of multimodal stimuli 10

1.2 Allocation of attention during processing of multimodal stimuli 13

1.3 The modality effect 15

1.4 Aims of this study 16

1.5 Utilizing eye data to infer implicit cognitive processes 18

1.6 Summary 19

Chapter 2 21

Multimodal attentional and perceptual processes: Differences in looking patterns 21

between congruity-incongruity manipulations of visual and auditory inputs 21

2.1 Introduction 21

2.2 Methods 25

2.2.1 Participants 25

2.2.3 Apparatus 26

2.2.2 Stimulus Materials 26
2.2.4 Experimental Design and Procedure 28
2.2.5 Statistical Analysis 30
2.3 Results 32
2.3.1 Dwell Time 32
2.3.2 Fixation Count 34
2.3.3 Congruity-Incongruity Ratings and Verbal Response 36
2.4 Discussion 38
2.5 Conclusion 41

Chapter 3 42
Multimodal recognition memory: Differences in pupillary response 42
between old/new manipulations of visual and auditory inputs 42
3.1 Introduction 42
3.2 Methods 45
3.2.1 Participants 45
3.2.3 Apparatus 46
3.2.2 Stimulus Materials 46
3.2.4 Procedure 47
3.2.5 Experimental Design and Statistical Analysis 51
3.2.6 Data Cleaning 52
3.3 Results 52
3.4 Discussion 54
3.5 Conclusion 57

Chapter 4 58

Multimodal imagery: Differences in spatial image generation 58

of visual and auditory cues 58

4.1 Introduction 58

4.2 Methods 62

4.2.1 Participants 62

4.2.4 Apparatus 62

4.2.2 Stimulus Materials 63

4.2.3 Procedure 65

4.2.5 Statistical Analysis 70

4.3 Results 71

4.3.1 Perceptual Encoding phase: Percentage of dwell time during stimulus presentation 71

4.3.2 Imagery phase: Percentage of dwell time 72

4.3.3 Analyses of incorrect trials 74

4.4 Discussion 74

4.5 Conclusion 77

Chapter 5 78

General Discussion 78

References 83

Appendix A - J ..... i - xii

List of Figures

Figure 1 Sample of the stimuli 27

Figure 2 The manipulation of the visual stimuli with regards to the visual context and the auditory stimuli 29

Figure 3 An example of the AOI definition in a stimulus..... 31

Figure 4 Graph showing the significant effect of the interaction between Congruency and Modality on mean dwell time in the Area of Interest (AOI). Error bars represent the standard errors of the mean..... 33

Figure 5 Graph showing the significant effect of the interaction between Congruency and Modality on mean fixation count in the Area of Interest (AOI). Error bars represent the standard errors of the mean..... 35

Figure 6 Graph showing the significant effect of the interaction between Congruency and Modality on mean fixation count outside the Area of Interest (AOI). Error bars represent the standard errors of the mean. 36

Figure 7 The congruity-incongruity ratings for each experimental condition. Bars represent the mean value of the congruity rating for each of the four kinds of context and sound manipulation (i.e., congruent context-congruent sound (CC), congruent context-incongruent sound (CI), incongruent context-congruent sound (IC), and incongruent context-incongruent sound (II)). The Y axis represents the congruity rating (i.e., 5 indicates high congruity rating and 1 indicates low congruity rating). Error bars represent the standard errors of the mean. 37

Figure 8 Percentage of participants considering the visual vs auditory incongruity as being more salient. 38

Figure 9 Illustration of the experimental procedure for all phases 48

Figure 10 Graph showing pupillary response (as deviation from baseline) across conditions. Error bars represent the standard error of the mean (SEM). 53

Figure 11 Graph showing the significant effect of the interaction between Novelty and Modality on pupillary response. Error bars represent the standard errors of the mean…….. 54

Figure 12 Illustration of the trial sequence in both experimental phases 69

Figure 13 Graph showing the mean percentage of dwell time in each quadrant during the Image Generation task, the Image Inspection task and the Vividness Rating task in the Imagery phase. Error bars represent the standard errors of the mean. The dotted grey line represents chance levels. 73

List of Tables

Table 1 Experimental phases and conditions..... 50

Chapter 1

General Background

Information processing usually involves the management and integration of multiple sensory channels (e.g., vision and audition), and it is therefore a complex process. Understanding how it works, however, is essential, because the information processing system filters relevant information and makes it available to make decisions and plan actions. To date, it is still unclear how multimodal information is exactly processed and integrated (Schneider, Engel, & Debener, 2008). Given that the environment we live in is obviously multimodal, this gap in our knowledge is especially problematic.

In this thesis, I will specifically investigate how multimodal information is processed, recognized and retrieved. In particular, I will study (i) how attention is allocated toward congruent and incongruent multimodal stimuli, (ii) how multimodal stimuli are recognized in recognition memory tasks, and (iii) how they are retrieved during spatial imagery activity.

1.1 Processing of multimodal stimuli

In order to explain how multimodal information is processed, three main different mechanisms have been proposed. Firstly, a more traditional view suggests that perceptual information is maintained exclusively in modality-specific perceptual systems, with the visual and auditory modalities being perceived as independent and separate units (Greene, Easton, & LaShell, 2001). As stimuli are separately processed

from an early stage, the multimodal integration would only take place during subsequent higher cognitive processing (Schneider et al., 2008). Secondly, other scholars have argued that processing multimodal stimuli flexibly relies on separate but interacting perceptual systems (Schneider et al., 2008). According to this hypothesis, the perceptual processing would occur in a modality-specific area, but the integration of this multimodal information would take place at an early stage (Andersen, Snyder, Bradley, & Xing, 1997). Finally, other authors have proposed that perceptual information is maintained in a combined representation system regardless of the input modality (Vandierendonck, 2016).

Although there is no clear consensus yet on how multimodal stimuli are exactly processed, it is clear that multimodal stimuli are processed differently than unimodal stimuli (e.g., stimuli in only one modality). However, psychologists disagree on the advantages of processing multimodal versus unimodal stimuli (Sinnett, Soto-Faraco, & Spence, 2008). On the one hand, multimodal stimuli would contribute to the richness of sensory experience (Diaconescu, Alain, & McIntosh, 2011), and would thus be more likely to be accurately detected and efficiently processed, as compared to unimodal stimuli (intersensory facilitation: Röder & Büchel, 2009; Tsilionis & Vatakis. 2016) . On the other hand, when humans are simultaneously presented with stimuli from different modalities, performance in one modality may thrive at the costs of the others (Dunifon et al., 2016), suggesting that multimodal processing may have a competitive nature (sensory competition: Sinnett et al., 2008).

Typically, behavioural studies have examined the effect of unimodal versus multimodal processing by investigating differences in cognitive efficiency (e.g., in terms of higher accuracy rate, lower response time, faster recognition and

identification; Lewandowski & Kobus, 1993; Molholm, Ritter, Javitt, & Foxe, 2004; Bahrick, Lickliter & Flom, 2004; Sinnett et al., 2008; Delogu, Raffone & Belardinelli, 2009). These studies have repeatedly shown that multimodal processing, by combining inputs from different modalities, is cognitively more efficient than unimodal processing (Dunifon, Rivera, & Robinson, 2016). Thompson and Paivio (1994), for example, presented participants with stimuli either in the dual-modality (picture-sound) or in the single-modality (picture-picture, sound-sound). When participants were asked to freely recall these stimuli, performance was better for dual- than single-modality stimuli. Crucially, mere within-modality repetitions (e.g., two pictures of the same object) were not sufficient to increase cognitive efficiency in a similar way, suggesting that it is the multimodal presentation of stimuli that helps to increase participants' ability to recall the stimuli.

Similarly, Goolkasian and Foos (2005) presented participants with word stimuli in different formats and modalities to test whether stimuli in the dual-modality (e.g., picture and spoken word, printed and spoken word) were processed more efficiently than stimuli in the single-modality (i.e., picture and printed word), in a working memory task. The results confirmed that stimuli in the dual-modality were better recalled than stimuli in the single-modality (i.e., with dual formats within the same modality). Furthermore, the pictures presented with the printed words were not recalled any better than either visual component alone (i.e., picture or printed word alone). These findings can be explained with Mayer's cognitive theory (Mayer & Anderson, 1991; Mayer & Sims, 1994), proposing that the visual-spatial sketchpad and the phonological loop systems of working memory are two separate but interconnected channels that process visual/pictorial versus auditory/verbal stimuli (see Baddeley, 2003 for a review). When stimuli are presented as pictures or as

spoken words, they are processed directly within one of these channels. However, as printed words involve visual and verbal components, they are processed in a more complex manner, with attention being split between the two channels. Therefore, although printed words are initially represented in the visual channel, this information is later transferred to the verbal channel for further processing. In contrast, when processing two formats of the same modality (e.g., visual formats), only one channel is used (e.g., the visual-spatial sketchpad). Information overloads the channel and causes participants to split their attention between formats rather than using the processing resources to build connections between visual and auditory channels, strengthening memory representation and reinforcing learning (Mayer & Sims, 1994; Moreno & Mayer, 2002). Therefore, stimuli in the dual-modality would be easier to process than stimuli in the single-modality with more formats.

1.2 Allocation of attention during processing of multimodal stimuli

The human brain continuously deals with a stream of complex sensory inputs from different modalities, which compete for visual awareness and control of action (Chun, 2000; Min, Zhai, Gao, Hu, & Yang, 2014). To find a way through this impressive amount of inputs and make informed decisions, humans rely on a cognitive control mechanism called attention, which can be considered a sort of cognitive filter. A primary role of attention is to selectively prioritize the processing of important sensory inputs from the environment, while discarding less important ones, thus avoiding cognitive overload (Summerfield & Egner, 2009; Talsma, 2015; Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010). Without the attentional mechanism,

humans would not be able to handle the tremendous amount of environmental inputs they are continuously exposed to.

Clearly, attentional allocation plays an important role to sort out and select the most relevant inputs also when processing multimodal stimuli. While processing multimodal information, attention can be selectively directed in different ways. For instance, attention can be directed to a specific modality (e.g., paying attention to auditory inputs while ignoring visual ones, or vice versa). Moreover, the focus of attention can be spatially-based (e.g., on a location in space), temporally-based (e.g., on a moment in time), or it can be based on the structural properties of the stimuli (e.g., the colour or size of a visual stimulus, or the pitch and loudness of a sound (Koelewijn, Bronkhorst, & Theeuwes, 2010).

Several theories have been proposed to explain how the attention system exactly works (Koelewijn et al., 2010; Mishra, 2015; Talsma et al., 2010). One of the most influential ones, the theory of biased competition by Desimone and Duncan (1995), claims that multiple stimuli compete for selection until attention focuses on one of them, and only the most salient stimuli are processed. According to this theory, attentional allocation can happen through bottom-up and top-down processes. The bottom-up (exogenous) process is stimulus driven, as stimuli involuntarily attract attention toward their salient properties (Beck & Kastner, 2009; Chun, 2000; Röder & Büchel, 2009). In contrast, the top-down (endogenous) process is voluntary, with individuals using cognition (e.g., prior knowledge, goals, instructions, memory, expectations, emotions, or expertise) to control the stimuli attended (Borji, 2014; Chen et al., 2014; Coco, Malcolm, & Keller, 2014; Koelewijn et al., 2010).

1.3 The modality effect

When humans are presented with stimuli in different modalities, their response to them may differ (i.e., modality effect; Colavita, 1974). In his seminal experiment, Colavita (1974) showed that the visual modality is dominant over the auditory modality in adults, as visual stimuli are processed more quickly and with higher accuracy than auditory stimuli. In his study, he randomly presented visual (light) or auditory (tone) stimuli to participants, who had to press one key for the visual stimulus and another key for the auditory stimulus. In few trials, stimuli with both auditory and visual modalities were presented. Surprisingly, in this dual-modality trials participant showed a tendency to only press the key for visual stimuli. After the experiment, some participants reported that they had failed to perceive the auditory stimulus in the bimodal trials, possibly as a result of an attentional bias favouring the visual modality (Colavita 1974). In line with this, Broadbent (1957) proposed that attention has a limited processing capacity, and it can only handle information from one modality at a time. Therefore, when there are simultaneous multimodal stimuli, the attentional system needs to switch from one modality to the other one, sequentially (Mishra, 2015; Sinnett, Spence, & Soto-Faraco, 2007).

Such a dominance of the visual modality, however, only emerges through development (Nava & Pavani, 2013). In infants and young children, the auditory modality is indeed dominant (Sloutsky & Napolitano, 2003; Sloutsky & Robinson, 2008). Using a similar experimental approach to the one by Colavita (1974), for instance, Nava and Pavani (2013) found that auditory dominance persists in children until 6 years of age, while the transition towards visual dominance starts in children aged 9 to 12. Similar developmental patterns were also found when using different

experimental procedures (see e.g., Shams, Kamitani, & Shimojo, 2002). These developmental changes in the modality dominance may reflect physiological processes: while the auditory system is responsive to external stimuli already before birth, the visual system only start being fully stimulated after birth (Nava & Pavani, 2013).

Moreover, such a dominance of the visual modality is only limited to certain contexts, because different modalities may have a different relevance in different contexts (Reinwein, 2012). In particular, certain sensory modality are processed more accurately only within their appropriate dimension (i.e., modality appropriateness hypothesis; Welch & Warren, 1980). Vision, for example, may be best suited for spatial processing tasks, while audition for temporal processing tasks (Lukas, 2009; Welch & Warren, 1980; Colavita, 1974; Talsma et al., 2010). Furthermore, human response to stimuli from different modalities dynamically changes depending not only on the stimuli used (Yuval-Greenberg & Deouell, 2009), but also on the context (Dunifon et al., 2016) and on the task demands (Sinnett et al., 2008).

1.4 Aims of this study

According to Mayer's cognitive theory (Mayer & Anderson, 1991; Mayer & Sims, 1994), processing multimodal stimuli is a complex phenomenon, with attention being split between channels. To date, it is still unclear how attention is exactly allocated between stimuli in different modalities, and how different modalities and context interplay on the allocation of attention when processing multimodal stimuli. The first aim of our study was therefore to disentangle the role of context and modality on the allocation of visual attention (see Chapter 2 for more details).

The second aim of our study was to understand whether the modality of the stimuli used affects recognition memory (i.e., the ability to identify old information and distinguish it from novel one; Kafkas & Montaldi , 2015; Võ et al., 2008). According to Colavita (1974), for instance, the visual modality is dominant over the auditory modality in adults. Therefore, visual stimuli are processed more efficiently than auditory ones (e.g., Thorpe et al., 1996), and they may also be more efficiently recognized (e.g., Ballas, 1993). In this study, we therefore aimed to assess participants' reaction to old versus novel stimuli presented in the visual and auditory modalities, to understand whether visual stimuli are also more easily recognized than auditory stimuli, and how recognition memory varies depending on the modality used (see Chapter 3 for more details).

Finally, the third aim of our study was to investigate how visual and auditory stimuli are mentally reconstructed during mental imagery (i.e., the process of reconstructing mental images in the absence of corresponding sensory stimulations; Lacey, 2013). In line with the modality appropriateness hypothesis (Welch & Warren, 1980), visual stimuli are processed better than auditory ones in spatial processing tasks (Lukas, 2009; Welch & Warren, 1980; Colavita, 1974; Talsma et al., 2010). However, we do not yet know whether this effect is also present during mental imagery, and how visual and auditory stimuli interplay while mentally reconstructing images (see Chapter 4 for more details). Given that we live in a multisensory world and we continuously receive sensory inputs from multiple modalities, understanding how multimodal stimuli are processed, recognized and retrieved appears crucial.

1.5 Utilizing eye data to infer implicit cognitive processes

In this study, we used eye trackers to determine how the manipulation of audio-visual stimuli affects eye movement patterns and pupillary responses. Eye-trackers are a non-invasive camera-based system which uses infrared illumination to illuminate the eyes. It determines the gaze position and pupillary response by continuously analysing the angle changes between the centre of the pupillary and corneal reflection (Brisson et al., 2013; Ryan, Hannula, & Cohen, 2007). The use of eye-trackers to study stimuli processing has increased rapidly in the last years, and eye-trackers are now widely available and easier to maneuver, partly because modern video-based eye trackers simplify the eye-tracking recording process (Irwin, 2004).

To date, it is well known that gaze behaviour can provide a direct insight into individuals' interests and intent (Yun, Peng, Samaras, & Zelinsky, 2013). Already in 1967, Yarbus provided evidence of this, by asking participants to search for a specific information in a painting. Participants' gaze behaviour and looking patterns followed both the physical properties of the painted scene, and the goals and interests of the participants, suggesting that looking patterns can be controlled by both bottom-up and top-down processes (Duchowski, 2002; Hoffman & Subramaniam, 1995). Since then, the use of eye-tracking techniques to study cognitive processes has steadily increased, due to their ability to both measure response to stimulus properties and participants' mental processes.

The eye-tracking techniques typically use two types of measures (i.e., temporal and spatial ones). Some of the frequently used eye-tracking measures are average fixation duration, proportion of time spent on each area of interest (AOI), fixation count, fixation count on each AOI, gaze duration mean on each AOI, and fixation rate

(count/s) (Lai et al., 2013). The studies reported in this dissertation utilized several eye movement parameters and also pupillary response. In Chapter 2, we investigated attentional and perceptual processes with congruent and incongruent multimodal stimuli, by measuring dwell time and fixation count at the area inside and outside AOI. In Chapter 3, we studied recognition memory of old and new multimodal stimuli, by examining pupillary response. In Chapter 4, we investigated the retrieval of multimodal stimuli during mental imagery, by measuring dwell time in different AOIs.

1.6 Summary

Humans live in a multimodal environment and continuously receive a flow of simultaneous sensory inputs from different channels. To avoid perceptual overload and selectively focus on a limited amount of these inputs, humans have evolved an active and efficient cognitive mechanism to sort out the sensory experience received across multiple sensory channels (e.g., visual, auditory, olfactory, gustatory, haptic, proprioception, etc.).

In this work, I aimed to contribute to the study of multimodal processing by investigating how different stimuli are processed, recognized and retrieved when they are presented across multiple modalities. Using the eye-tracking method, I assessed how the experimental manipulation of audio-visual stimuli affects eye movements and pupillary behaviour during attentional and perceptual processes (Chapter 2), recognition memory (Chapter 3) and imagery (Chapter 4).

This dissertation is organized in five chapters. In the first chapter, I have provided the general theoretical background to the current experiments, especially focusing on the processing of stimuli across different modalities. In the second, third, and fourth chapters I will describe in detail the three studies I have conducted. Finally, the fifth chapter will present a general discussion on the main findings of these studies, with possible directions for future research.

Chapter 2

Multimodal attentional and perceptual processes: Differences in looking patterns between congruity-incongruity manipulations of visual and auditory inputs

2.1 Introduction

Typically, processing information requires the management and integration of multiple sensory channels. However, it is still largely debated how information from different sensory channels is integrated during multimodal processing (Schneider et al., 2008). Several studies, for instance, show that multimodal information allows individuals to better detect and identify target objects, as compared to information from only one modality (Colonius & Diederich, 2006; Fort, Delpuech, Pernier, & Giard, 2002; Giard & Peronnet, 1999; Miller, 1982; Molholm et al., 2004; Sinnett et al., 2008). However, when processing multimodal stimuli, performance in one modality may thrive at the costs of the others (Dunifon et al., 2016). In particular, several studies have shown that, when multimodal information is incongruent (i.e., if it fails to reflect regular associations between e.g., an object and a sound), adults maintain the same ability to process the dominant modality, but the ability to process the non-dominant modality may decrease (e.g., Colavita, 1974; Lewkowicz, 1988a, 1988b; Robinson & Sloutsky, 2004; Sloutsky & Napolitano, 2003). For example, when incongruent visual and auditory information is presented, adults are generally faster and more accurate to process visual than auditory stimuli (e.g., Colavita, 1974; Yuval-Greenberg & Deouell, 2009; Talsma, 2010).

To date, however, most studies on the effect of congruity on multimodal processing have failed to consider the manipulation of contextual information (Chen et al., 2014; Chen & Spence, 2010; Min et al., 2014; Suied, Bonneel, & Viaud-

Delmon, 2009; Vogler & Titchener, 2011; Yuval-Greenberg & Deouell, 2009). This is especially problematic, as human response to stimuli from different modalities dynamically changes also depending on the context in which they are provided (Freides, 1974). For instance, a policeman investigating a burglary case might hear a metallic clicking sound and wonder if it is a gun, while the same person sitting at home on the sofa would likely not associate the clicking sound to a gun. This example from Ballas and Mullins (1991) demonstrates how the context can influence sound identification.

In contrast, several studies on stimuli processing have systematically manipulated the relation between target object and context, but they have mostly failed to use a multimodal approach, including for instance no auditory information during manipulations (Coco et al., 2014; Davenport, 2007; Davenport & Potter, 2004; Fiedler, 2013; LaPointe, Lupianez, & Milliken, 2013; Mudrik, Deouell, & Lamy, 2011; Ralph, Seli, Cheng, Solman, & Smilek, 2014; Underwood & Foulsham, 2006; Underwood, Templeman, Lamming, & Foulsham, 2008; Võ & Henderson, 2011).

So far, only few studies have manipulated both multimodal and contextual information (Laurienti, Kraft, Maldjian, Burdette, & Wallace, 2004; Özcan & van Egmond, 2009). For example, using a contextual priming paradigm, Özcan and van Egmond (2009) studied the effect of visual context on the identification of environmental sounds (i.e., air, alarm, cyclic, impact, liquid, mechanical sounds). They found that visual context positively affected the identification of ambiguous environmental sounds. However, the degree of the contextual effects depended on the physical and semantic character of the sound (e.g., alarm sounds were inherently identified better and faster than other sound types, whereas impact sounds had

inherently shorter durations and were more difficult to identify). Even though this study investigated the interplay of visual context and sound, it failed to use a variety of living and non-living auditory stimuli, and also failed to investigate the effect of congruity manipulations of context and sound. Moreover, most previous studies relied on behavioural measures such as accuracy rates and response time to infer individuals' performance, with only few studies using more objective eye-tracking data to investigate these issues (Chen et al., 2014; Min et al., 2014). However, monitoring eye movements with eye-tracking techniques is especially useful, as it allows more reliably following participants' allocation of attention and cognitive processing (Mishra, 2015; Zelinsky, 2013).

Given these premises, it is clear that little is still known about how different modalities and context interplay on the allocation of attention during the perceptual processing of congruent and incongruent multimodal stimuli. In this study, we therefore aimed to investigate how the multimodal (i.e., visual-auditory) presentation of different stimuli affected visual attention (i.e., selectivity in one's visual field; Cohen, 2013) through congruity and incongruity manipulations of context and multimodal stimuli. The target object (e.g., a chicken) could be either congruent/incongruent with the context (e.g., a farm versus a living room), and/or with an auditory stimulus (e.g., a chicken sound versus a cat sound).

To date, the effect of incongruences between target objects and context on looking patterns is still unclear. Võ and Henderson (2011), for instance, investigated the influence of object-context inconsistencies on eye movement control, when observing pictures. They found that participants did not show preferential gaze towards the regions of inconsistency, likely because the object-context inconsistency

weakened contextual guidance, impeding search performance and efficient eye movement control. Similarly, Coco et al. (2014) found no evidence of longer looking duration when objects and context were incongruent. In particular, visual attention preferentially focused on contextually congruent objects rather than contextually incongruent objects, especially if objects were visually salient. Coco et al. (2014) provided several explanations for their results. Firstly, contextually congruent objects would compete for attentional resources, reducing looking duration on incongruent objects. Being semantically irregular, in contrast, incongruent objects can be more easily remembered and are thus less dependent on attentional processing mechanisms. Secondly, following the cognitive relevance framework, contextually congruent objects may be processed before incongruent objects, because incongruent objects do not fit the top-down representational knowledge and/or the contextual expectations.

Although these studies provide evidence that incongruity may not attract attention, other authors have argued that attention is indeed preferentially directed toward incongruent stimuli or events, as compared to congruent ones (e.g., Henderson, 1992). This hypothesis is based on the schema hypothesis, according to which individuals develop expectations about objects, based on the memory representation of prototypical scenes (Henderson, 1992). Thus, the violation of such perceptual expectations or schema is expected to attract individual attention more than a congruent, expected event (Ralph et al., 2014; Underwood & Foulsham, 2006; Underwood et al., 2008). According to this hypothesis, we therefore predicted that participants would overall look longer when target object and context are incongruent, as compared to when they are congruent.

Finally, we also expected general differences in looking patterns depending on the modality of the stimuli used. Auditory stimuli, for instance, are generally considered to be more alerting than stimuli in other modalities (Posner, Nissen, & Klein, 1976), and they are also processed more slowly (Ballas, 1993; Brunetti, Indraccolo, Mastroberardino, Spence, & Santangelo, 2017; Viggiano et al., 2017) than visual stimuli, which are instantaneously processed (Chen & Spence, 2011). Auditory stimuli may thus elicit greater attention than visual stimuli, because of their greater saliency and longer processing time. Therefore, we predicted that looking time would be overall longer for auditory than visual stimuli. Given that, to our knowledge, no previous study has analyzed the complex interaction of context, congruity and modality, we made no detailed predictions about how attention would be exactly allocated between context and target object, depending on the congruity and modality of the stimuli used.

2.2 Methods

2.2.1 Participants

Thirty-four individuals (29 females, 5 males) participated in the study (mean age = 23.18, *SD* = 4.41). All participants gave informed consent according to the guidelines of the University of Bern institutional ethics review board. All participants reported normal or corrected-to-normal vision and hearing. They were all naïve with regards to the purpose of the experiment conducted. Participants received a course credit in return for their participation.

2.2.3 Apparatus

Eye data were recorded using a video-based iView X RED tracking system (SensoMotoric Instruments, Teltow, Germany) integrated with a 17-inch TFT monitor. The system captured the eye data with a sampling rate of 50/60 Hz, a tracking resolution of < 0.01º, and a gaze position accuracy of < 0.5º (SensoMotoric Instruments, 2009). This infrared remote eye-tracking device was contact-free, and the system allowed the automatic compensation of head movements by tracking the corneal reflex. Participants were calibrated using a 5-point calibration with validation. Stimulus presentation and data collection were controlled using SMI Experiment Center software (SensoMotoric Instruments, Teltow, Germany). Eye-tracking data were extracted using Be-Gaze software (SensoMotoric Instruments, Teltow, Germany).

2.2.2 Stimulus Materials

The multimodal stimuli were created by incorporating the image and audio components. Auditory stimuli were gained from various internet sources. Five raters were appointed to rate the recognizability of each audio file. Only the sounds with high ratings were utilized as stimuli. Visual stimuli (object and context image) were obtained from Google image and Flickr. The static images were used in form of 2D photographs. Each set of stimuli consisted of the target object in the contextual background, accompanied by an audio source. The context and audio were either congruent or incongruent with the target object, producing four different types of stimuli in each set: congruent context-congruent sound (CC), congruent context-incongruent sound (CI), incongruent context-congruent sound (IC), and incongruent

context-incongruent sound (II). A sample of stimuli-set is illustrated in Figure 1, and a complete list of the stimuli-sets is provided in the Appendix.

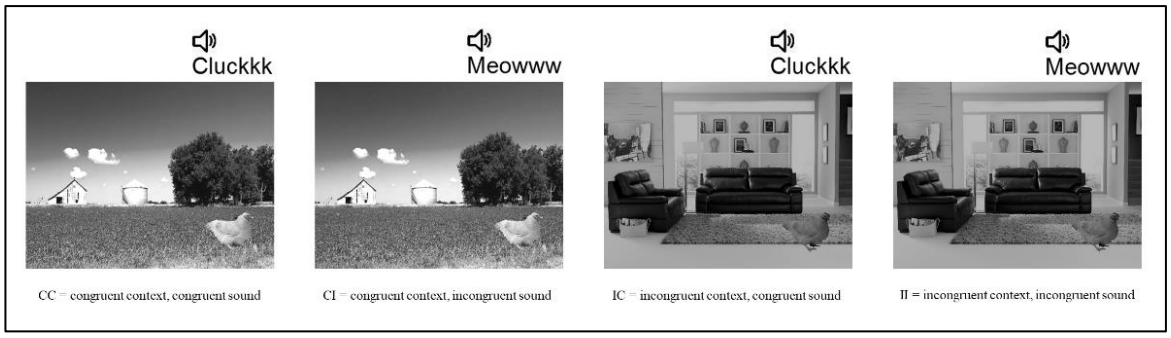


Figure 1 Sample of the stimuli

To create congruent and incongruent stimuli, photos were manipulated using the Paint.net and Inkscape Software. The dimension of each photo was set to 640 x 480 pixels, and each target object integrated in the context image was set to 200 x 200 pixels. SHINE (spectrum, histogram, and intensity normalization and equalization) toolbox in MATLAB was used to control the low-level image properties (Willenbockel et al., 2010).

Each experimental condition consisted of 30 stimuli, including various animate and inanimate objects. Context images were relatively complex but contained no single items that could be falsely perceived as target objects (e.g., the size of the target object was always larger than any other object in the context). Moreover, the position of the target object in each image was controlled for, dividing each context image into four quadrants and placing the target object in a different quadrant for each set of stimuli.

2.2.4 Experimental Design and Procedure

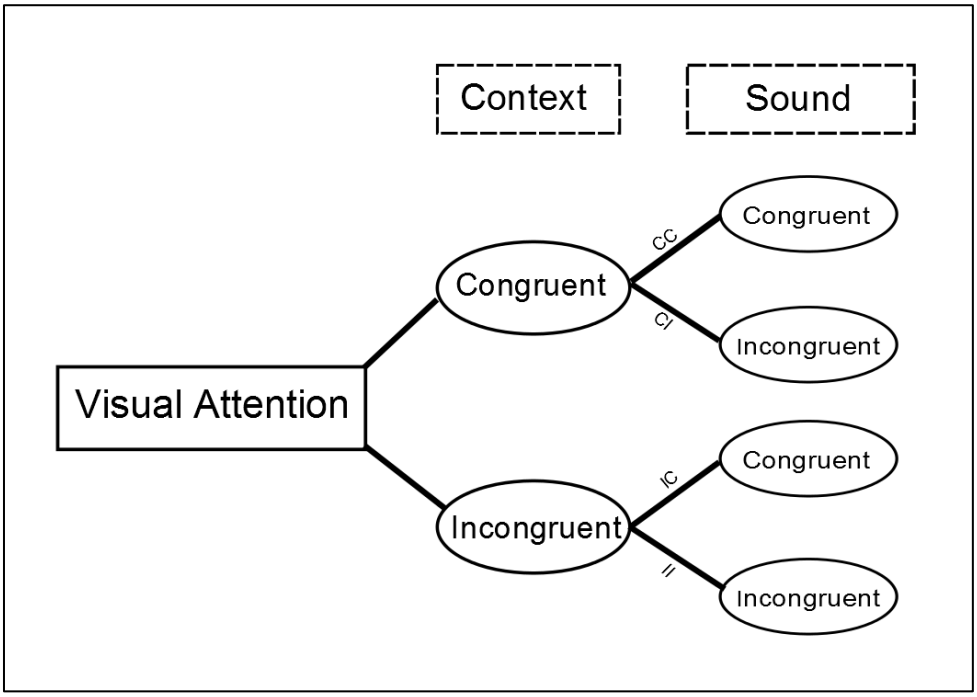
Participants were tested individually in a dimly lighted and quiet sound-attenuated eye-tracking laboratory. They sat at a distance of approximately 60-70 cm from the monitor screen. The distance varied slightly because participants were free to move their head and body. Participants were instructed to look at the screen during the presentation of the stimuli.

Trials started with a five-point calibration and validation method. The participants' position was adjusted until allowing us to accurately collect eye movements within 0.8 visual degrees. Before each stimulus presentation, a fixation cross was displayed at the centre of the screen for 1500 ms. The audio and visual stimuli were then simultaneously presented for 5000 ms, the audio content being delivered through a headphone.

All participants underwent a pre-test session, a test session, and a rating session. A pre-test session was conducted to train participants and make them familiar with the eye-tracking experimental setting and procedure. The pre-test session consisted of five trials, with 3 congruent stimuli and 2 incongruent stimuli, which were not used in other sessions.

The stimuli presented during a test session belonged to one of four different conditions, in which the context and/or the sound of the stimulus could be manipulated (Figure 2). The four conditions were: congruent context-congruent sound (CC), congruent context-incongruent sound (CI), incongruent context-congruent sound (IC), and incongruent context-incongruent sound (II). The order of the conditions (CC, CI, IC, and II) was pseudorandomized.

In the rating session, we used the same stimuli as in the test session. Participants were required to explicitly rate the congruity-incongruity level of each stimulus. The question “How coherent is the stimulus?” appeared on the screen, and participants had to choose one of five-responses (e.g., 1 was “Not coherent at all”, 5 was “Very coherent”), by selecting an answer with a mouse click. At the end of the experimental session, participants were verbally asked to name the most evident incongruities, to obtain an explicit response about the modality effect.



* CC=congruent context, congruent sound; CI=congruent context, incongruent sound; IC=incongruent context, congruent sound; II=incongruent context, incongruent sound

Figure 2 The manipulation of the visual stimuli with regards to the visual context and the auditory stimuli

2.2.5 Statistical Analysis

The analyses were conducted to examine the potential differences between multimodal conditions in looking behaviour. Eye movements were analyzed in the test sessions from the onset of the stimulus until its offset (5000 ms). The analyses were based on fixations and calculated using Be-Gaze software, SensoMotoric Instruments. Fixations were detected when the sum of the dispersion of the gaze stream on the x and y axes was below 100 pixels and when the duration exceeded 80 ms. The Area of Interest (AOI) for each stimulus in each condition was the target object. The AOIs were defined by drawing a square around the target object as illustrated in Figure 3.

The mean dwell time (i.e., the time duration of one visit in an AOI, from entry to exit; Holmqvist et al., 2011) and the mean fixation count (i.e., the number of fixations in an AOI; Holmqvist et al., 2011) were calculated for both the area inside the AOI (target object) and outside the AOI (context). These two parameters provided information about looking patterns, indicating the extent of attentional allocation towards the stimulus presented (i.e., as a measure of attention and active searching behaviour, respectively). The mean dwell time and mean fixation count of looking at the area in AOI and also at the area out of AOI were compared with repeated measures ANOVA with four levels (CC, CI, IC, II), to assess variation in looking patterns depending on the congruity of the stimuli in different modality sources. Additionally, we conducted a repeated measure 2 (Congruency: Congruity, Incongruity) × 2 (Modality: Visual, Auditory) ANOVA, to understand the specific effect of congruity-incongruity manipulation of visual and auditory inputs on the mean dwell time looking at the target object. To accomplish this analysis, the mean dwell time of conditions with Congruent Visual input (CC and CI), Congruent

Auditory input (CC and IC), Incongruent Visual input (IC and II), and Incongruent Auditory input (CI and II) were calculated for each participant. For all tests, partial eta-squared (η^2_p) are reported as a measure of effect size. A Huynh-Feldt correction was applied to the degrees of freedom of those tests when the assumption of sphericity was violated. The alpha level for all the statistical tests was set at .05. When effects were significant, we conducted post-hoc comparisons, using Tukey adjustments to correct for multiple comparisons. In the Results, we presented all significant post-hoc comparisons, but in the Discussion we only focused on those comparisons in which only one parameter differed, as differences were easier to interpret (e.g., we included comparisons of congruent auditory stimuli vs incongruent auditory stimuli, or congruent visual stimuli vs congruent auditory stimuli, but not congruent auditory stimuli vs incongruent visual stimuli). Finally, we coded participants' subjective congruity-incongruity ratings and verbal responses, by calculating the individual averages for the responses given.

Figure 3 An example of the AOI definition in a stimulus

2.3 Results

2.3.1 Dwell Time

The Congruency x Modality ANOVA revealed a significant main effect of Congruency ($F(1, 33) = 18.90$, $p < .001$, $\eta^2_p = .36$), and a significant effect of the interaction between Congruency × Modality ($F(1, 33) = 5.91$, $p < .05$, $\eta^2_p = .15$), but no significant main effect of Modality ($F(1, 33) = 0.00$, $p > .05$), on looking time in the AOI (i.e., of the target object) . The results are shown in Figure 4. Post-hoc analyses revealed that the mean dwell time looking at the target object was significantly longer when the stimuli contained congruent visual inputs (408.52 ms) rather than incongruent auditory inputs (379.21 ms). Moreover, participants looked significantly longer at the target object when the stimuli contained congruent auditory inputs (431.26 ms) rather than incongruent visual inputs (401.96 ms) or incongruent auditory inputs (379.21 ms).

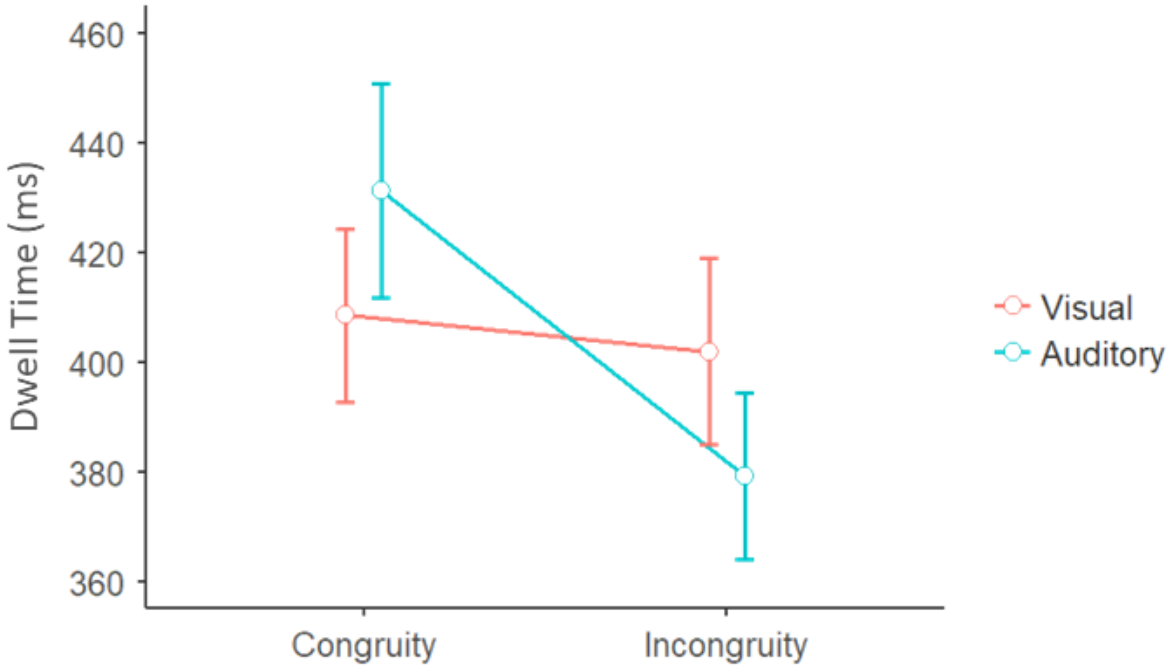


Figure 4 Graph showing the significant effect of the interaction between Congruency and Modality on mean dwell time in the Area of Interest (AOI). Error bars represent the standard errors of the mean.

A similar 2 × 2 ANOVA revealed a significant main effect of Congruency ($F(1, 33) = 4.66$, $p < .05$, $\eta^2_p = .12$) on looking time outside the AOI (i.e., of the context). There was no significant effect of Modality ($F(1, 33) = 0.00$, $p > .05$, $\eta^2_p = .00$) and no significant effect of the interaction Congruency × Modality ($F(1, 33) = 1.40$, $p > .05$, $\eta^2_p = .04$), on looking time outside the AOI. Post-hoc analyses did not confirm significance.

2.3.2 Fixation Count

Similarly, the Congruency x Modality ANOVA revealed a significant main effect of Congruency ($F(1, 33) = 59.5$, $p < .001$, $\eta^2_p = .64$), and a significant effect of the interaction between Congruency × Modality ($F(1, 33) = 84.6$, $p < .001$, $\eta^2_p = .72$), but no significant main effect of Modality ($F(1, 33) = 0.00$, $p > .05$, $\eta^2_p = .00$) on fixation counts in the AOI. The results are shown in Figure 5. Post-hoc analyses revealed that mean fixation count was significantly higher when the stimulus contained congruent auditory inputs (5.74) rather than incongruent auditory inputs (4.53), congruent visual inputs (5.10) or incongruent visual inputs (5.17). Similarly, mean fixation count was significantly higher when the stimulus contained congruent visual inputs (5.10) rather than incongruent auditory inputs (4.53), and when the stimulus contained incongruent visual inputs rather than incongruent auditory inputs.

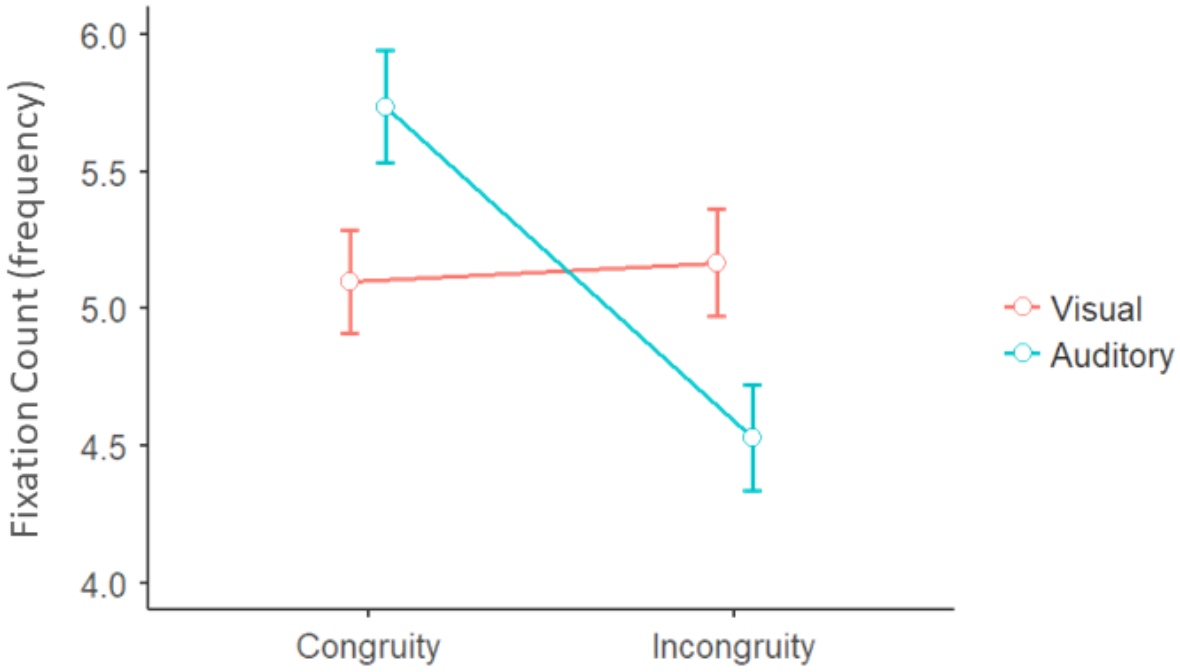


Figure 5 Graph showing the significant effect of the interaction between Congruency and Modality on mean fixation count in the Area of Interest (AOI). Error bars represent the standard errors of the mean.

A similar 2 × 2 ANOVA revealed a significant main effect of Congruency ($F(1, 33) = 83.3$, $p < .001$, $\eta^2_p = .72$), and a significant effect of the interaction between Congruency × Modality ($F(1, 33) = 58.9$, $p < .001$, $\eta^2_p = .64$), on fixation count outside the AOI (i.e., of the context), but no significant main effect of Modality ($F(1, 33) = 0.00$, $p > .05$, $\eta^2_p = .00$). Post-hoc analyses revealed that fixation count outside the AOI was significantly higher with stimuli containing incongruent auditory inputs (9.79) rather than congruent auditory inputs (7.53), congruent visual inputs (8.65) or incongruent visual inputs (8.68). Fixation count was also significantly higher when stimuli contained congruent visual inputs (8.65) or incongruent visual inputs (8.68) rather than congruent auditory inputs (7.53). The results are shown in Figure 6.

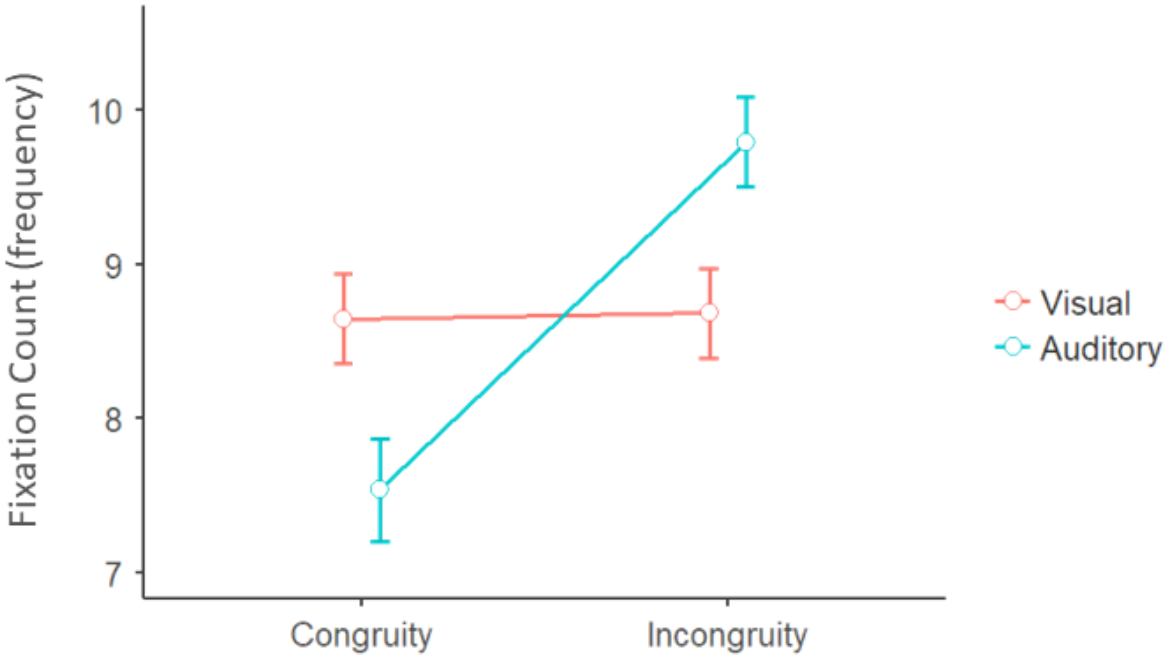


Figure 6 Graph showing the significant effect of the interaction between Congruency and Modality on mean fixation count outside the Area of Interest (AOI). Error bars represent the standard errors of the mean.

2.3.3 Congruity-Incongruity Ratings and Verbal Response

Participants rated stimuli with congruent sound (i.e., CC and IC) as being high in congruity (CC: 4.91, *SD* 0.29; IC: 3.23 *SD* 0.09). On the contrary, stimuli with incongruent sound (i.e., CI and II) were rated low in congruity (CI: 2.58, *SD* 0.15; II: 1.4, *SD* 1.39). The results are shown in Figure 7.

In addition, the majority of participants (*n* = 18, 52.9 %) found that auditory incongruity was more salient. Thirteen participants (38.2 %) found visual incongruity to be more salient, while the other participants could not decide (*n* = 2) or gave no answer (*n* = 1). This suggests that decisions on perceived congruity-incongruity were

made based on the auditory source component rather than on the visual source component. The results are shown in Figure 8.

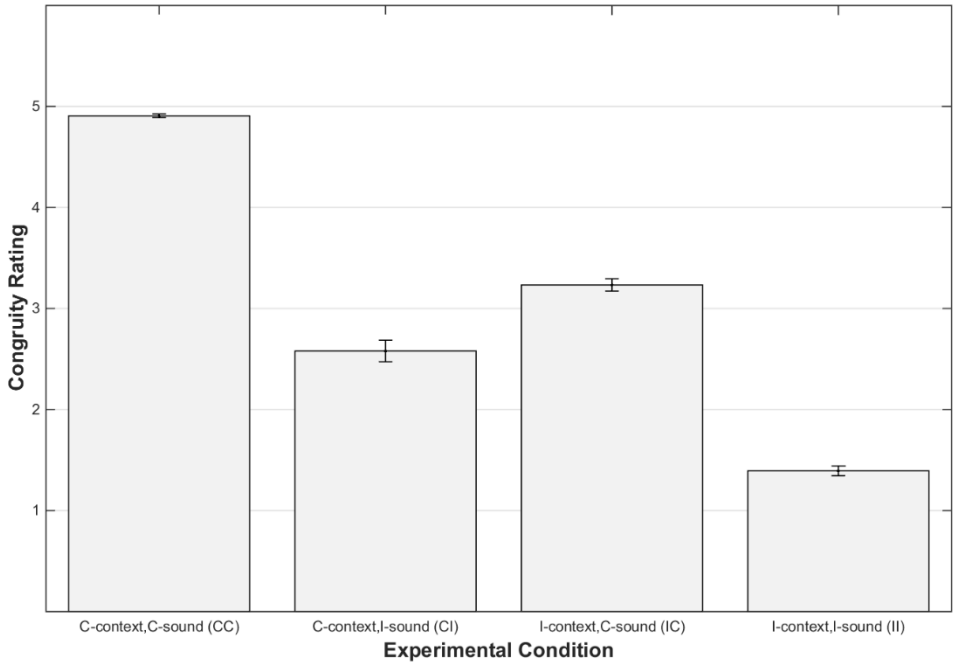


Figure 7 The congruity-incongruity ratings for each experimental condition. Bars represent the mean value of the congruity rating for each of the four kinds of context and sound manipulation (i.e., congruent context-congruent sound (CC), congruent context-incongruent sound (CI), incongruent context-congruent sound (IC), and incongruent context-incongruent sound (II)). The Y axis represents the congruity rating (i.e., 5 indicates high congruity rating and 1 indicates low congruity rating). Error bars represent the standard errors of the mean.

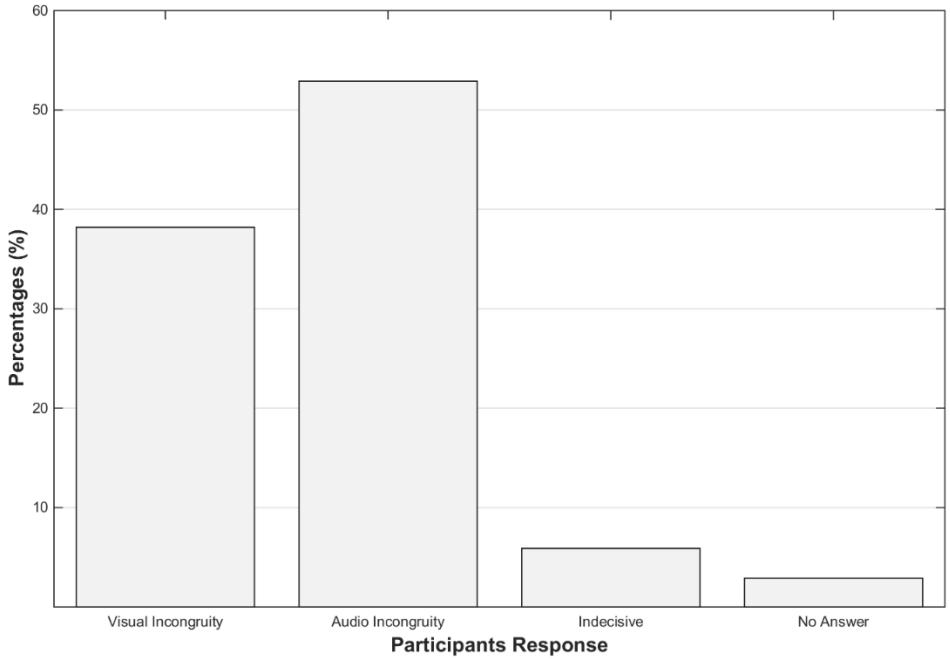


Figure 8 Percentage of participants considering the visual vs auditory incongruity as being more salient.

2.4 Discussion

In this study, we used eye-tracking measures to investigate how different modalities and context interplay on the allocation of visual attention during the perceptual processing of congruent and incongruent multimodal stimuli. Our results showed significant differences in looking patterns across manipulations. In particular, subjects allocated more visual attention to the target object (i) when auditory stimuli were congruent (as compared to when they were incongruent), (ii) to auditory stimuli (rather than visual ones) when stimuli were congruent, and (iii) to visual stimuli (rather than auditory ones) when stimuli were incongruent. Exactly the opposite pattern was evidenced for the allocation of visual attention to the context, which was higher (i) when auditory stimuli were incongruent (as compared to when they were congruent), (ii) to visual stimuli (rather than auditory ones) when stimuli were

congruent, and (iii) to auditory stimuli (rather than visual ones) when stimuli were incongruent.

Overall, this study revealed a complex interaction of congruity, modality and context, which affected the way participants allocated their attention. In particular, participants allocated their attention differently between target object and context, with these differences being modulated by both the congruity and the modality of the stimuli. These results are important, because they suggest that attention allocation is a very complex phenomenon, and that the interaction of multiple factors (e.g., context, stimulus modality) should be better taken into account when designing this kind of studies. Moreover, our results may help explaining contradicting findings of previous studies, as participants' response may strongly differ even if little procedural changes are introduced.

This study showed that participants' attention preferentially focused on the target object when auditory stimuli were congruent, and on the context when auditory stimuli were incongruent. Possibly, when auditory stimuli are incongruent (e.g., when hearing a cat sound, while a chicken is visually displayed), participants may react to the incongruency by looking for an alternative plausible auditory source in the context (e.g., scanning the context in search of a cat). Therefore, looking time would be longer outside the AOI when auditory stimuli are incongruent.

Our findings also showed that participants preferentially allocated attention to the context, when visual stimuli were congruent, and to the target object, when visual stimuli were incongruent. When the context and the target object are incongruent, participants may preferentially focus their attention to the object, because the object is the only "unconnected" item in an otherwise homogeneous group of items (i.e., the

context). In other words, participants would preferentially allocate their attention toward the target object, to try and solve the incongruency by actively searching for example, yet overlooked characteristics of the target object. During perception, indeed, humans use ‘scene schema’, ‘schemata’ or ‘context frames’ that contain conceptual knowledge about the environment. This contextual structure is viewed as a set of expectations that can facilitate perceptual experience by guiding the acquisition of information (Bar, 2004; Chun, 2000). While objects that are congruent with the schema are more easily and reliably processed (Davenport & Potter, 2004; Stubblefield, Jacobs, Kim, & Goolkasian, 2013), incongruent objects may be harder to process, leading to longer looking times.

This study had several limitations. Firstly, we only used two measures (i.e., dwell time and fixation count) to assess how participants allocated their attention. Future studies, instead, should investigate how context, modality and congruity affect attention in humans by using a different response mode, or eye-tracking measures. In addition, it may be interesting to use other measurement methods, such as pupillometry or neuroimaging. Pupillary response, for instance, has been considered as a reliable index of arousal and implicit cognitive processing (Sirois & Brisson, 2014), while neuroimaging methods may explore how different brain areas are affected by modality and congruity manipulations. Finally, future studies may use modalities other than the visual and auditory ones to investigate attention allocation in humans, following it through development.

2.5 Conclusion

Experiment 1 showed that adults allocate their visual attention differently, depending on the context, congruency and modality of the stimuli used. Our results show that these factors interplay in a complex way, and that incongruities between visual and auditory inputs produce different attentional and perceptual experiences, which result in different attention allocation between target objects and contexts. These findings are limited by the nature of the audio-visual stimulus employed, and generalization to broader categories of participants and stimuli can only be determined by further experiments.

Chapter 3

Multimodal recognition memory: Differences in pupillary response between old/new manipulations of visual and auditory inputs

3.1 Introduction

Recognition memory involves the ability to identify old information and distinguish it from novel one (Kafkas & Montaldi, 2015; Võ et al., 2008). This is something that we continuously do in our everyday life, for instance when we meet people at social events, and we quickly have to recall whether we know them already. Clearly, recognition memory is essential to ensure our normal social and non-social functioning, by allowing us to reliably recognise familiar people and objects. Therefore, investigating the cognitive mechanisms related to recognition memory is crucial to understand how human brain works when processing old information and integrating it with new one.

In typical recognition memory tasks, participants are presented with a set of old stimuli (i.e., already observed) and new stimuli (i.e., not yet observed in the task). Then, participants are asked which stimuli have been already observed and which ones are novel (for a review, see Yonelinas, 2002). In these tasks, participants can recognize old stimuli with two different mechanisms: recollection or familiarity. Recollection is a form of explicit memory, which happens when participants consciously identify the specific details of the item or the contextual information available. In contrast, familiarity is a form of implicit memory happening without conscious awareness: participants have "the feeling" of having a memory of the stimulus, but make no explicit association with its contextual details (Kafkas &

Montaldi, 2015; Küper, Groh-Bordin, Zimmer, & Ecker, 2012; Yonelinas, 2002). Although it is not always clear which of these two mechanisms is used to recognize old stimuli, both probably play an important role in recognition memory.

Research on recognition memory has mostly used event-related potential (ERP) and pupillometry as psychophysiological measures to analyse how humans react when confronted with old versus novel stimuli (Brocher & Graf, 2017). Previous studies using ERP, for instance, have found that old stimuli and new stimuli elicit different ERP waveform contours at different points in time and at different locations over the scalp (Brocher & Graf, 2016). For example, different components (i.e., FN400 and P600, or late positive component, LPC) are reliably used to detect participants' response to old stimuli (Curran & Friedman, 2004; Küper et al., 2012; Voss & Paller, 2008). Similarly, pupil dilation can provide reliable information on memory retrieval, arousal, emotion, and cognitive effort (Kahneman & Beatty, 1966; Kahneman & Peavler, 1969; Kloosterman et al., 2015; Naber, Frässle, Rutishauser, & Einhäuser, 2013), and it is thus a highly sensitive marker of memory processing (Gomes, Montaldi, & Mayes, 2015; Papesh, Goldinger, & Hout, 2012). Pupil dilation, for instance, is greater towards older than new stimuli in recognition memory tests, possibly because different neural and cognitive mechanisms are involved in the recognition of old stimuli (Kafkas & Montaldi, 2015, 2017; Montefinese, Ambrosini, Fairfield, & Mammarella, 2013). In particular, recognizing old stimuli would require the conscious retrieval of associative information from the encoding event and thus posit a higher cognitive load (Kafkas & Montaldi, 2017; Rugg & Curran, 2007; Võ et al., 2008), which is in turn linked to a higher pupil dilation (see e.g., van der Wel & van Steenbergen, 2018).

To date, several studies have investigated recognition memory of visual stimuli. To our knowledge, however, no experimental studies on recognition memory have yet been carried out across modalities. Would participants also differ in their reaction to old versus novel stimuli (i.e., old/new effect), when these were presented in two distinct modalities? And would there be differences in the old/new effect, depending on the modality used? Understanding how multimodal stimuli are recognized is especially important, as our environment is fundamentally multimodal and, in our everyday life, stimuli recognition usually happens across different modalities.

Moreover, recognizing multimodal stimuli is especially complex, as stimuli in different modalities are processed differently (Dunifon et al., 2016), and perception of a stimulus in one modality is affected by perception of another stimulus in a different modality (i.e., inter-sensory bias; Lukas et al., 2010). Visual stimuli may be more efficiently processed and thus more quickly recognized than auditory ones (e.g., Colavita, 1974). Thorpe, Fize, and Marlot (1996), for example, have shown that participants are able to detect and recognize the presence of a wide range of animals integrated in complex visual scenes within less than 150 ms after stimulus onset, confirming the great processing efficiency of the visual system in object perception and recognition. In contrast, sounds require more time to be identified, and this may slow down the recognition process (e.g., Ballas, 1993).

In this study, we therefore presented participants with images of various geometrical shapes accompanied by different sounds. Participants were required to learn and memorize each shape-sound association, and they were later tested in a recognition memory task. Based on previous findings on the old/new effect (i.e.,

greater dilation for old stimuli, as compared to new stimuli), and on the notion that visual stimuli are usually processed more quickly than auditory stimuli (see above), we predicted that participants (i) would demonstrate larger pupil dilation when presented with old auditory as compared to novel auditory stimuli (as it happens in the visual modality; e.g., Kafkas & Montaldi, 2017; Rugg & Curran, 2007; Võ et al., 2008), and (ii) would overall demonstrate larger pupil dilation with auditory than visual stimuli. In particular, pupillary dilation should be highest for stimuli with old sounds, and lowest for stimuli with novel shapes.

3.2 Methods

3.2.1 Participants

Forty-six individuals (36 females, 10 males) participated in the study (mean age = 21.11, *SD* = 2.07). All participants gave informed consent according to the guidelines of the University of Bern institutional ethics review board. All participants reported normal or corrected-to-normal vision and hearing. They were all naïve with respect to the purpose of the experiment conducted. Participants received a course credit in return for their participation. Data points from two participants were excluded from data analysis due to the low measurement values of tracking ratio (both participants had a tracking ratio below 80%, ranging from 48.5% to 64.2%). Data points from trials showing inconsistent and inaccurate responses (as described below) were also excluded from further analysis.

3.2.3 Apparatus

We used the same apparatus described in Chapter 2. To record pupillary response, we used a video-based iView X RED tracking system (SensoMotoric Instruments, Teltow, Germany), integrated with a 17-inch TFT monitor. This infrared remote eye-tracking device was contact-free and allowed the automatic compensation for head movements by tracking the corneal reflex. Before collecting pupillary response data for each participant, a calibration procedure for the eye tracker was performed using a 5-point calibration and validation method (Ramdane-Cherif & NaÏt-AliNait-Ali, 2008).

3.2.2 Stimulus Materials

As stimuli, we used 40 pictures of 2-D symmetrical shapes which were paired with 40 sounds (i.e., 20 animate and 20 inanimate sounds). The 40 stimuli were divided into two stimuli set (i.e., Set A and Set B).

Visual stimulus: We used 40 different grey-coloured symmetrical shapes (20 as old and 20 as novel stimuli). The dimension of the shapes was approximately 5 × 5, 5 × 7, or 7 × 5 cm (width × height). The shapes were presented at the centre of the screen against a white-coloured rectangular background measuring 33 × 27 cm.

Auditory stimulus: We used 40 different animate and inanimate sounds (20 as old stimuli and 20 as novel ones). The sounds were obtained from internet. These auditory stimuli were presented via headphones at a comfortable hearing level.

To create multimodal stimuli, each shape was paired with a 5000 ms recorded sound of an animate or inanimate object (i.e., vehicles, musical instruments and household items). In order to avoid multimodal stimuli that could suggest a semantic meaning, we used pairs of meaningless audio-visual stimuli, by for instance pairing the shape of a triangle with the sound of a cat, or a square with the sound of a car. Associations with no semantic meaning were necessary (i) to avoid that participants associated a label to the stimuli, and (ii) to reduce the probability of interference during the encoding processes of the stimuli, in case of pre-existing knowledge (Thelen, Talsma, & Murray, 2015).

3.2.4 Procedure

All participants were tested individually in a dimly lit eye-tracking laboratory. They sat at a distance of approximately 70 cm from the monitor screen. Prior to the actual experiment, the participants did a Pre-test to ensure they were familiar with the eye-tracking experimental setting and procedure. After the Pre-test, the experiment started. It consisted of three phases of testing: a learning phase, a recall phase and a test phase. Figure 9 shows the experimental procedure for all phases.

Each phase began with a calibration and validation procedure, in which participants had to look to the screen and move their eyes to follow some dots moving on the screen. Then, participants were briefly reminded on the experimental procedures. Each trial started with a fixation screen (2000 ms), and participants looking at the central fixation point. Then, the audio-visual stimuli were presented for 5000 ms.

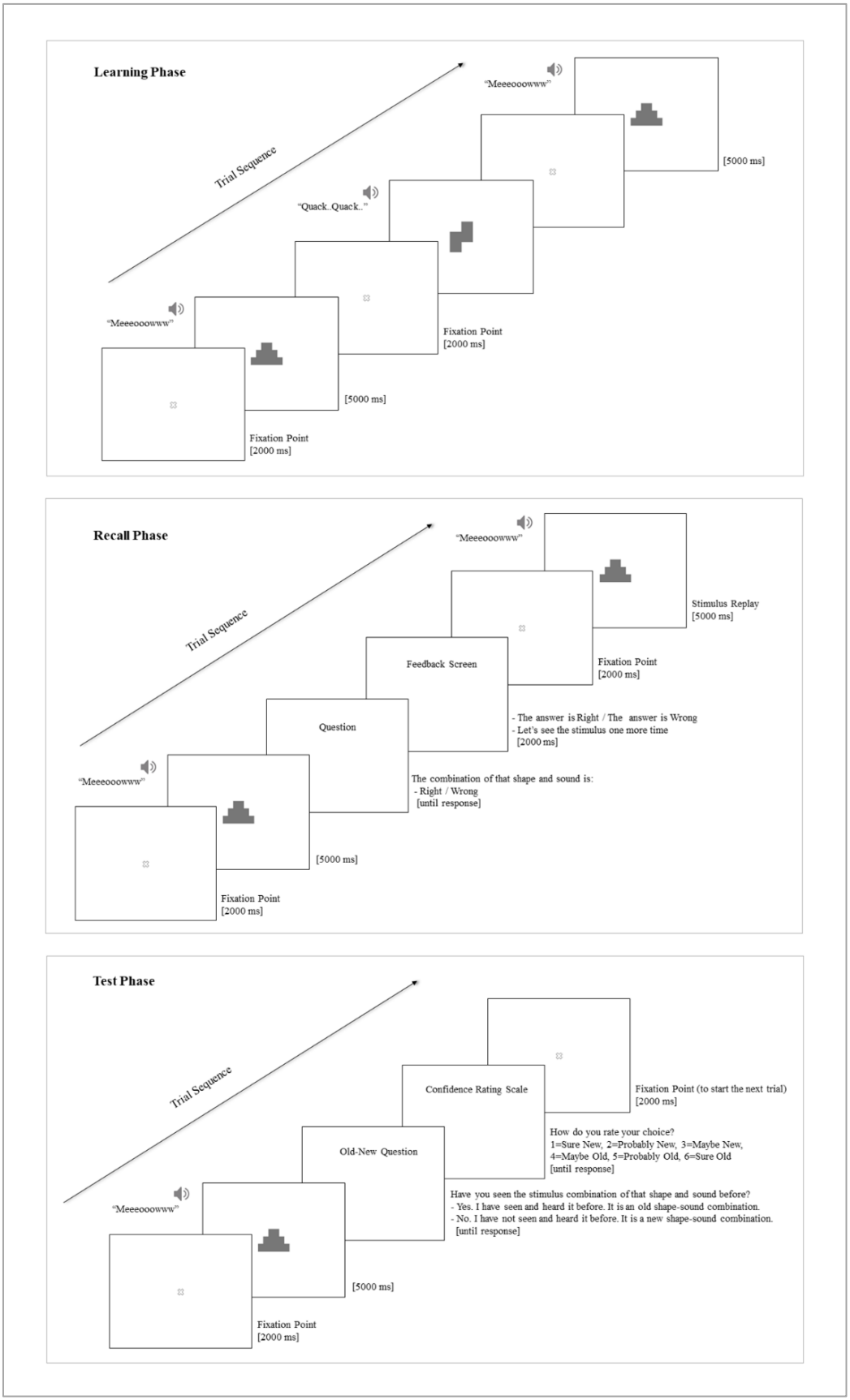


Figure 9 Illustration of the experimental procedure for all phases

Learning phase: During the learning phase, participants were presented an encoding task in which they had to learn and remember new associations between shapes and sounds. Participants were shown ten stimuli of shape-sound pairs, and each pair was presented three times, randomizing the order of the pairs.

Recall phase: During the recall phase, participants were tested in a discrimination task, in which they needed to identify whether the shape-sound pair was right or wrong (i.e., like the one shown in the learning phase, or not). Half of the stimuli presented during the Recall phase were correct, and the other half was incorrect. The participants' answer was followed by a feedback screen displaying whether the shape-sound pair was correct or incorrect. The purpose of the Recall phase was to strengthen the memory traces for the studied stimuli. Ensuring that participants reliably recognize the stimuli was crucial to compare pupil dilation with old and new stimuli during the next phase.

Test phase: During the Test phase, participants were presented with the old and new stimuli and were subsequently asked to make a recognition judgement. We presented participants with four different conditions, each characterized by the use of different multimodal stimuli: i) old shape - old sound, ii) old shape - new sound, iii) new shape - old sound, and iv) new shape - new sound. No mention was made to participants about the different types of stimuli presented. Table 1 illustrates the stimuli presented in each Phase.

After each stimulus presentation, participants were asked to determine whether the stimulus was an old or new one, by showing them the following question on the screen: '*Have you seen the stimulus combination of that shape and sound before?*'. Participants then had to select one of the two response options: '*Yes. I have seen and*

heard it before. It is an old shape-sound combination', and '*No. I have not seen it before. It is a new shape-sound combination*'. Right after, the following question appeared on the screen: '*How do you rate your choice?*'. Participants could thus provide a recognisability rating using a six-point Likert scale (*1=Surely new, 2=Probably new, 3=Maybe new, 4=Maybe old, 5=Probably old,* and *6=Surely old*; see Wixted (2009). The confidence rating method has been widely used to study recognition memory (Hales & Brewer, 2011; Papesh et al., 2012).

Participants gave all their responses by clicking on the selected response with a mouse. After the participant had chosen a response, a new trial began. The experiment lasted for approximately 30 minutes. After the experiment, participants were debriefed about the purpose of the experiment. Course credits were granted for their participation.

Table 1 Experimental phases and conditions

Phase	No. of Stimulus	Condition
Learning	10 x 3	- each stimulus is presented three times
Recall	20	- 10 stimuli with the right shape - sound combination - 10 stimuli with the wrong shape - sound combination
Test	40	- 10 stimuli with old shape - old sound - 10 stimuli with old shape - new sound - 10 stimuli with new shape - old sound - 10 stimuli with new shape - new sound

3.2.5 Experimental Design and Statistical Analysis

In the Results section, we only present the data collected during the Test phase. As dependent variable we used the pupillary response, which consisted in the difference (in number of pixels) between the average pupil size during stimulus presentation and during baseline. Pupillary response during stimulus presentation was computed from the moment the stimulus was shown, until it disappeared. Pupillary response in the baseline was computed as the average pupil size during the fixation screen. The baseline correction was made to minimize trial-to-trial fluctuations in the pupillary signal (Brocher & Graf, 2016). Pupillary response was calculated for each participant and trial.

The aim of the analyses was to explore how pupillary response varied depending on the novelty of the multimodal stimuli. The first analysis was a one-way repeated measures analysis of variance (ANOVA) on pupillary response, to determine whether there were statistically significant differences across experimental conditions. We used Tukey adjustments to correct for multiple comparisons, reporting effect size as partial eta squared. Huynh-Feldt correction to the degrees of freedom was used when the sphericity assumption was violated. The alpha level for statistical tests was set at .05. The second analysis was a repeated measures of 2 (novelty: Old, New) × 2 (modality: Shape, Sound) ANOVA, to test how pupillary response varied depending on the interaction between novelty and modality. As dependent variable, we used the individual mean pupillary response to old shape, old sound, new shape and new sound (calculated for e.g., old shape by averaging responses in the old shape - old sound and in the old shape - new sound conditions, and so on).

3.2.6 Data Cleaning

To prepare data for the analyses, we used three criteria. Firstly, we excluded data for all the participants who had a tracking ratio below 80% in the Test phase. As a result, data from two participants were omitted. Secondly, we also excluded all data points containing inconsistent responses (i.e., participants assessed the stimuli as being old, but in the subsequent question rating recognisability they assessed the stimuli as being novel). As a result, 14 data points (i.e., 0.4 % of the whole data set) were deleted. Thirdly, we excluded all data points containing inaccurate responses (i.e., participants assessed as novel the old shape - old sound combination, or they assessed as old the other three combinations). As a result, 57 data points (i.e., 1.64 % of the whole data set) were deleted.

3.3 Results

Pupillary Response Analysis. A one-way repeated-measures ANOVA with the old/new multimodal manipulation as within-subject factor showed that there was a statistically significant difference across conditions ($F(3,129) = 5.85$, $p < .001$, $\eta^2_p = .12$). Post-hoc tests revealed that the mean pupillary response in the condition old shape - new sound was significantly higher (0.61, $SD = 0.49$) than in the new shape - old sound condition (0.38, $SD = 0.40$). No other comparison was statistically significant. Figure 10 illustrates pupillary response in each condition.

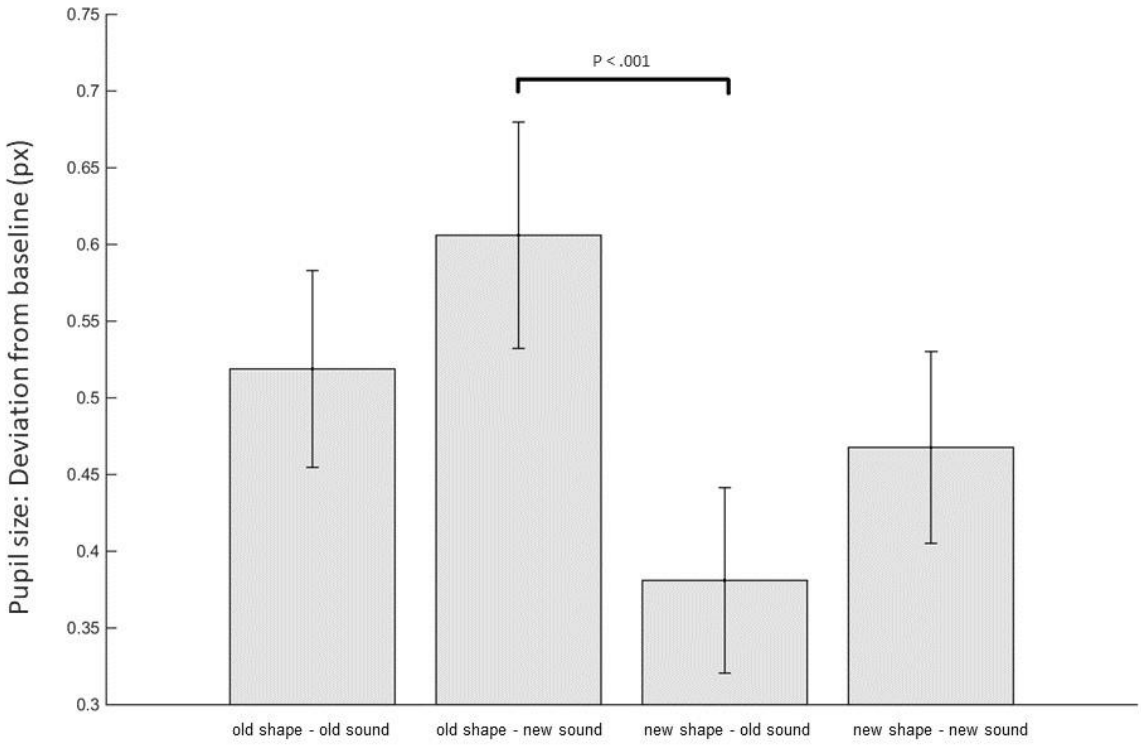


Figure 10 Graph showing pupillary response (as deviation from baseline) across conditions. Error bars represent the standard error of the mean (SEM).

A repeated measures of 2 (novelty: Old, New) × 2 (modality: Shape, Sound) ANOVA revealed a significant interaction between novelty and modality ($F(1, 43) = 12.79$, $p < .001$, $\eta^2_p = .23$). No significant main effect of novelty ($F(1, 43) = 1.04$, $p > .05$, $\eta^2_p = .02$) and modality were found ($F(1, 43) = .003$, $p > .05$, $\eta^2_p = .00$). The results are shown in Figure 11. Post-hoc analyses revealed that pupils dilated significantly more when participants were presented with 'Old-Shape' stimuli (0.56 px) than both 'Old-Sound' stimuli (0.45 px) and 'New-Shape' stimuli (0.42 px). Moreover, pupils dilated more when participants were presented with 'New-Sound' stimuli (0.54 px) than 'New-Shape' stimuli (0.42 px).

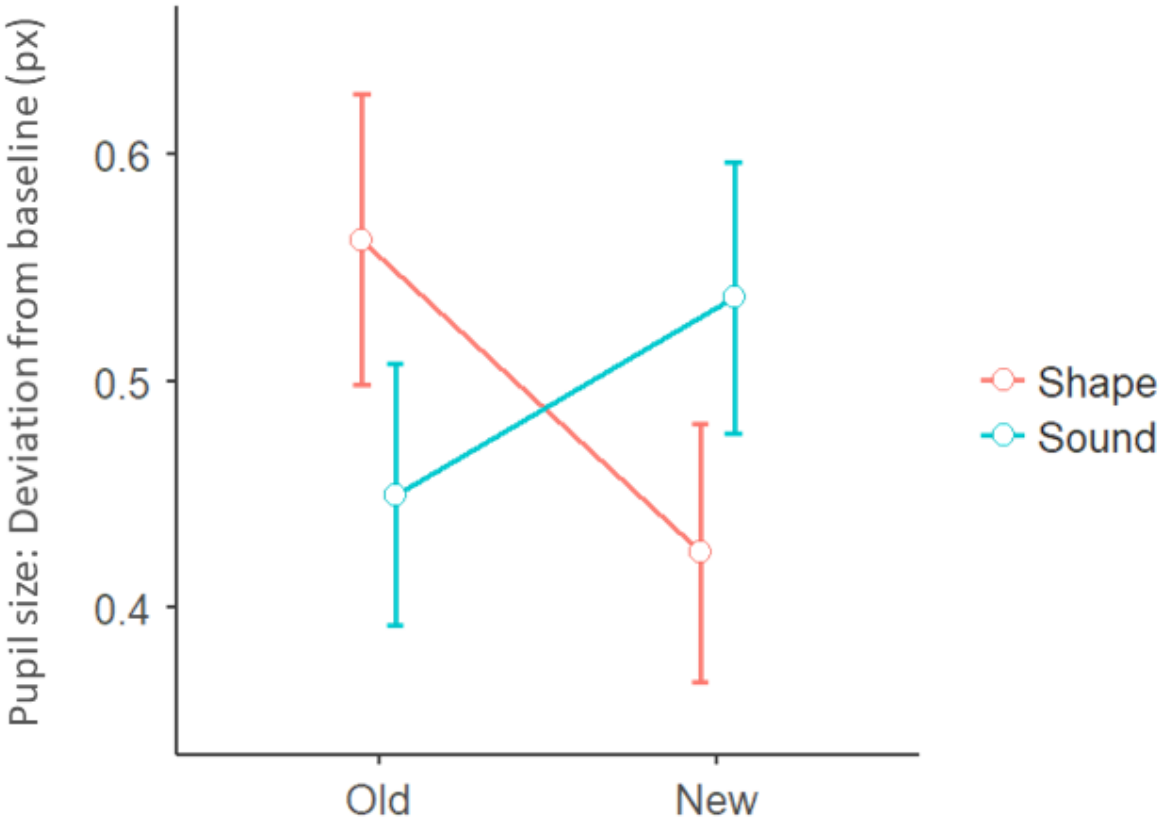


Figure 11 Graph showing the significant effect of the interaction between Novelty and Modality on pupillary response. Error bars represent the standard errors of the mean.

3.4 Discussion

In this study, we investigated recognition memory of multimodal stimuli, focusing on the participants' reaction to old versus novel stimuli (i.e., old/new effect) presented in the visual and auditory modalities. Our results revealed significant differences in recognition memory, depending on the novelty and modality of the stimuli. Firstly, pupillary response significantly differed only in the old/new mismatched pairs, with old shape - new sound stimuli eliciting a higher pupillary response than new shape - old sound stimuli. Secondly, pupillary response was

significantly higher with old visual stimuli (as compared to old auditory stimuli and novel visual stimuli), and also with novel auditory stimuli (as compared to novel visual stimuli).

Pupillary response was significantly higher when old visual stimuli were paired with novel auditory stimuli, as compared to novel visual stimuli paired with old auditory ones. Thus, the interplay of the stimuli familiarity and modality clearly produced different effects on pupillary response. Interestingly, these differences were not significant when comparing stimuli which only differed along one dimension (i.e., familiarity *or* modality).

Why should old shape - new sound stimuli therefore be so hard to process? According to our predictions, old auditory stimuli should be the hardest ones to process, while novel visual stimuli should be the easiest ones, but this was not the case. One reason why old shape - new sound stimuli elicited the highest pupillary response may be that processing old shapes is especially demanding. Surely, old stimuli are harder to process than novel ones, because they imply recollection processes in recognition memory (e.g., Colavita, 1974; Thorpe, et al., 1996; Ballas, 1993), but visual stimuli are also notoriously easier to process than auditory ones (e.g., Colavita, 1974; Thorpe, et al., 1996; Ballas, 1993). Therefore, this is not a likely explanation of our results. Another reason why pupillary response was highest in old shape - new sound stimuli may be that new auditory stimuli are especially hard to process, although no memory recognition is involved. In particular, processing novel sounds may require higher cognitive effort and cognitive control (Botvinick, Braver, Barch, & Carter, 2001), leaving little cognitive resources to the recognition of shapes. Possibly, novel sounds are especially relevant for humans, and largely monopolize

their attention (SanMiguel, Linden, & Escera, 2010). Therefore, the higher pupillary response in old shape - new sound stimuli would largely depend on the higher cognitive demands of processing novel sounds. This would also explain why novel auditory stimuli were in general as hard to process as old ones, in contrast with the old/new effect (see below).

To also more generally assess the effect of modality and novelty independently of each other, we further run a 2 x 2 ANOVA. However, these results should be taken with caution, because the effect of modality and novelty on pupillary response is indeed an interaction. That said, our results showed that older stimuli were overall harder to process than novel ones, but only in the visual modality (i.e., pupillary response was higher in old shape than in new shape), because in the auditory modality no significant effect was found (i.e., pupillary response did not differ between old sound and new sound). The fact that older stimuli were harder to process than novel ones is in line with abundant literature on recognition memory, and confirms that more complex neural and cognitive mechanisms are involved in the recognition of old stimuli (e.g., Kafkas & Montaldi, 2015, 2017; Montefinese, Ambrosini, Fairfield, & Mammarella, 2013; Rugg & Curran, 2007; Võ et al., 2008). However, our results were also partially unexpected, in that this old/new effect did not extend to the auditory modality. As briefly explained above, these results may suggest that the saliency of novel auditory stimuli is so high for humans that the perception of auditory stimuli primarily draws upon the available cognitive resources. Therefore, auditory stimuli would be harder to process when they are novel, and not when they are old, in contrast with the effect of recognition memory, which is known for visual stimuli. These results are interesting, and will need to be validated by more experiments in the future.

Our results also partially confirmed previous results from literature, showing that auditory stimuli are harder to process than visual stimuli (e.g., Colavita, 1974; Thorpe, et al., 1996; Ballas, 1993). In our study, however, this was only true for the novel stimuli (i.e., pupillary response was higher with new sound than with new shape), but not for the old ones (i.e., pupillary response was higher with old shape than with old sound). Why should old visual stimuli be harder to process in our study? Possibly, these results were simply biased by the fact that, in some trials, old shape was paired with novel sounds: if novel sounds really elicit higher pupillary response, also old shapes might appear to be hard to process, although they are not. Therefore, these results need to be taken with caution.

3.5 Conclusion

This study investigated recognition memory of multimodal stimuli. We measured how participants' pupillary response differed, when they were presented with old versus novel stimuli in the visual and auditory modalities. Our Experiment showed that novelty and modality interplayed during recognition memory, with novel auditory stimuli eliciting the highest pupillary response. Our results challenge the common view that old stimuli are generally harder to process. Further investigation should confirm and expand on these results, integrating a multimodal approach to the study of recognition memory.

Chapter 4

Multimodal imagery: Differences in spatial image generation of visual and auditory cues

4.1 Introduction

In this study we used an eye-tracking approach to investigate how visual and auditory stimuli affect looking behaviour during spatial imagery activity. Imagery has been defined as a conscious sensory experience, in which mental images are reconstructed in the absence of actual corresponding sensory stimulations (Lacey & Lawson, 2013). According to literature, there are two different subsystems of imagery, i.e., object imagery and spatial imagery. Object imagery refers to the information processing related to the external appearance of objects and scenes in terms of colour, brightness, size, shape, and texture. In contrast, spatial imagery refers to the information processing related to the location of objects in space, spatial relationships between objects or parts of objects, movements of objects and object parts, and other spatial transformations such as mental rotation (Blajenkova, Kozhevnikov, & Motes, 2006; Johansson, Holsanova, & Homqvist, 2011).

Typically, imagery activity is considered to rely on similar motoric processes (i.e., eye movements) to the ones used during perception, with oculomotor experiences happening gradually and sequentially during both imagery and perceptual activity (Brandt & Stark, 1997; Laeng, Bloem, D'Ascenzo, & Tommasi, 2014). Hebb (1968), for instance, pointed out that when forming an image of a familiar object (such as a car), the internal representation of that image is not immediately clear, but it is sequentially integrated and organized. Similar processes happen when we look at

actual objects during perception, when we make a series of fixations at different parts of the object. Therefore, this scanning process is similar when we observe external stimuli during perception, and when we generate and inspect internal mental images during imagery. During both activities, the eye fixation falls sequentially from one part to the other of an image, producing a series of eye movements (Laeng & Teodorescu, 2002).

The function of these eye movements during imagery has long been unclear, and two main hypotheses have been put forward to try and explain it. According to the epiphenomenal account, eye movements during imagery play a passive role, and are the by-product of mental image generation processes, simply mirroring the internal scanning of an image (Richardson & Spivey, 2000). More recently, however, other scholars have argued that eye movements during imagery have an active functional role, facilitating the process of information retrieval and image generation (e.g., Laeng et al., 2014; Laeng & Teodorescu, 2002). This is because eye fixations during perception of the external stimuli would be stored along with their visual representation. Therefore, this information would be used as a spatial index in a motor-based coordinate system to properly arrange all the component parts of the mental image during imagery (Laeng et al., 2014; Laeng & Teodorescu, 2002).

Previous studies on imagery have investigated how information retrieval during imagery is facilitated by the re-enactment of these sequences of fixations acquired during perceptual encoding processes (Bochynska & Laeng, 2015; Laeng & Teodorescu, 2002; Richardson & Spivey, 2000). Mast and Kosslyn (2002), for example, argued that the higher the resemblance of scan-paths between perception and imagery episodes, the better participants performed in a subsequent spatial memory

task. Furthermore, Martarelli and colleagues showed that, during imagery tasks, participants looked longer at the areas where the stimuli had been previously encoded (i.e., 'corresponding area effect'; Martarelli, Chiquet, Laeng, & Mast, 2017; Martarelli & Mast, 2013; Martarelli & Mast, 2011; Wantz, Martarelli, & Mast, 2015). This effect is argued to be robust and stable over time, as it can be detected also one week after perception, and can persist across different categories of items (Martarelli, et al., 2017; Wantz, et al., 2015). Therefore, these studies overall suggest a functional role and a non-rigid nature of eye movements during imagery.

Although several studies now converge in suggesting that eye movements during perceptual and imagery phases share a similar scan-path pattern (e.g., Martarelli et al., 2017; Bochynska & Laeng, 2015), little is still known on how visual and auditory inputs interplay during eye movements in imagery. Given that we live in a multisensory world and we continuously receive sensory inputs from multiple modalities, it is essential to also study imagery using ecologically more valid multimodal stimuli. A common assumption in sensory processing literature is that when visual and auditory stimuli are presented simultaneously, performance between modalities will differ: performance in one modality will thrive (i.e., modality dominance), while performance in the other modality will be hindered (Dunifon et al., 2016). As suggested by the modality appropriateness hypothesis, this modality effect is largely influenced by the contextual circumstances and natural characteristics of the stimuli (Freides, 1974; Welch & Warren, 1980). In particular, this model suggests that different sensory mechanisms are built upon unique structural properties, and each one is more suitable for specific tasks. In other words, the processing of stimuli in a certain sensory modality is more efficient within its appropriate dimension. For

example, vision may be best suited for spatial processing tasks, while audition for temporal processing tasks (Lukas, 2009; Lukas et al., 2010; Welch & Warren, 1980).

In this study we aimed to understand how visual and auditory stimuli are mentally reconstructed during imagery activity. For this reason, we administered participants with a spatial imagery task, and examined how eye fixation varied depending on the modality of the stimuli. We used pairs of stimuli (e.g., the image of a cat that was paired with the sound of a car, and the image of a car that was paired with the sound of a cat) in which the image appeared at different locations on the screen, and then we presented the same sound (e.g., the sound of a cat) to assess whether participants during imagery showed a longer fixation time at the quadrant which had been cued with a visual (a cat image) or an auditory (a cat sound) stimulus. The imagery task we presented was essentially a spatial task, in that we manipulated the spatial location of the visual and auditory stimuli presented. In line with the modality appropriateness hypothesis, we therefore hypothesized that participants would show longer dwell time where visual (rather than auditory) stimuli had been previously shown, since the visual component has a higher spatial resolution and is dominant in the processing of spatial characteristics (Talsma et al., 2010).

Additionally, we also aimed to investigate the effect of semantic category manipulations on gaze behaviour during imagery. While congruent stimuli facilitate processing in recognition tasks, incongruent stimuli hinder processing and lead to conflicts (Vogler & Titchener, 2011; Yuval-Greenberg & Deouell, 2009). However, some incongruencies may be stronger than others. In this study, we therefore used intra-categorical (e.g., animal image paired with animal sound) and extra-categorical (e.g., animal image paired with object sound) incongruent stimuli. Given that

processing of an object also depends on its semantic nature (e.g., living vs. non-living object; Viggiano et al. (2017), we expected incongruency to be stronger for the extra-categorical stimuli. In particular, we predicted that participants presented with intra-categorical stimuli would experience less conflict. Therefore, they should more easily recall the location of the visual stimuli, as compared to participants presented with extra-categorical stimuli, and should show longer dwell time in the area where the image had previously appeared.

4.2 Methods

4.2.1 Participants

Thirty individuals (24 females, 6 males) participated in the study (mean age = 21.40, *SD* = 1.79). All participants gave informed consent according to the guidelines of the University of Bern institutional ethics review board. All participants reported normal or corrected-to-normal vision and hearing. None reported a colour vision deficiency. They were all naïve in respect to the purpose of the study. Participants received a course credit in return for their participation. The experimental procedure took approximately 90 minutes to complete.

4.2.4 Apparatus

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

and a gaze position accuracy of < 0.5º (SensoMotoric Instruments, 2009). Other details on the eye-tracking system can be found in the previous chapters.

4.2.2 Stimulus Materials

We used different stimuli for each experimental phase (i.e., Pre-test, Encoding phase, and Imagery phase), as described below:

4.2.2.1 Pre-test

The Pre-test session consisted of five trials, in which we used non-meaningful audio-visual stimuli, mimicking the two following experimental phases. The images used for the Pre-test were chocolate, cheese, computer, river, and door. The stimulus presented in the Pre-test session were not presented again in the other phases.

4.2.2.2 Encoding phase

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

The auditory stimuli were the sounds usually associated to the 32 living animals and 32 non-living objects. The sounds were obtained from different internet sources, edited and saved as MPEG Layer 3, Stereo format at a sampling rate of 44,100 Hz. Each sound was 5000 ms in duration. The sound was presented over closed-ear headphones. The selection of the animal images and sounds mostly followed Viggiano et al. (2017), except for goose, rooster and turkey (i.e., 9.38 % of the animal stimuli used). Similarly, the selection of the images and sounds of non-living objects was taken from Viggiano et al. (2017), except for bicycle, fire truck, ship, bagpipes, drum, guitar, tambourine, alarm clock, lawn mower and printer (i.e., 31.25 % of the non-living object stimuli used).

In order to generate the audio-visual stimuli to be presented during the Encoding task, each image and sound from the stimulus pool were randomly paired. Collectively, there were 32 visual-auditory stimulus pairs. To study the influence of sensory modality on eye fixation during imagery activity, two variants of image-sound pairs were created from the pool of 32 bimodal stimulus items. Both these two variants included the same two stimuli (e.g., car and cat), but they were presented in different modalities (e.g., as image or sound), and images were located at different quadrants. For example, an image-sound pair of ‘cat-car’ could be either presented as a cat image paired with a car engine sound (with the cat image being located in e.g., the upper-left quadrant), or as a car image paired with a cat sound (with the car image being located in e.g., the bottom-right quadrant).

4.2.2.3 Imagery Phase

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

4.2.3 Procedure

Participants were tested individually in a dimly lighted and quiet sound-attenuated eye-tracking laboratory. They sat at a distance of approximately 60-70 cm from the monitor screen. The distance varied slightly because of the participants' freedom to move their head and body. Experimental session was preceded with a calibration procedure for the eye-tracker using a 5-point calibration and validation method during which participants had to fixate alternatively at the black calibration dots. Before each stimulus presentation, a fixation cross was displayed at the centre of the screen for 1000 ms.

4.2.3.1 Pre-test

A Pre-test session was conducted to train participants and ensure that they were familiar with the eye-tracking experimental settings and procedures. This session also ensured that the participants knew what to expect during the Encoding phase, and to verify that they understood how to perform each task during the Imagery phase.

4.2.3.2 Encoding Phase

During the Encoding phase, the audio-visual stimuli were presented simultaneously for 5000 ms. Stimuli were presented in a random order, to ensure that participants could not predict the position of any stimulus. Note that all the bimodal stimuli presented during the Encoding phase were semantically incongruent, as in Viggiano et al. (2017). There were two types of incongruent audio-visual manipulations: extra-categorical manipulations (e.g., the image of a cat paired with the sound of a car engine) and intra-categorical (e.g., the image of a tiger paired with the sound of a monkey). Half of the stimuli were extra-categorical and half were intra-categorical. By using these two incongruent categorical groups, we could assess whether distinct semantic representations affect gaze preference towards a particular quadrant during imagery task. Participants were randomly assigned to either the intra-categorical (Group A) or extra-categorical (Group B) conditions. The selection of animals and objects to be included in Group A and Group B was done by taking into consideration the physical characteristics and sounds/vocalizations produced by each animal or object, avoiding the use of similar animals or objects in the same group (e.g., if Group A contained a tiger and a goat, the lion and the sheep were assigned to Group B).

4.2.3.3 Imagery Phase

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

i) Image Generation Task.

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

ii) Image Inspection Task.

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

iii) Vividness Rating Task.

In the Vividness Rating task, participants heard the following statement: 'How do you rate the vividness of your mental image? Rate from 1 to 5. One is no visualization at all and five is perfectly clear'. After participants had given their response, the experimenter instantly pressed the space bar to proceed to the next screen.

It should be noted that the experimenter recorded the participants' response manually, and that each task in the Imagery phase was self-paced (i.e., there were no restrictions on speed and duration to accomplish the task). At the end of each trial, a final 'Stop' screen informed participants that they could eliminate their mental image. On the 'Stop' screen, participants were allowed to blink or rest their eyes before the next trial began. Figure 12 is an illustration of the trial sequence.

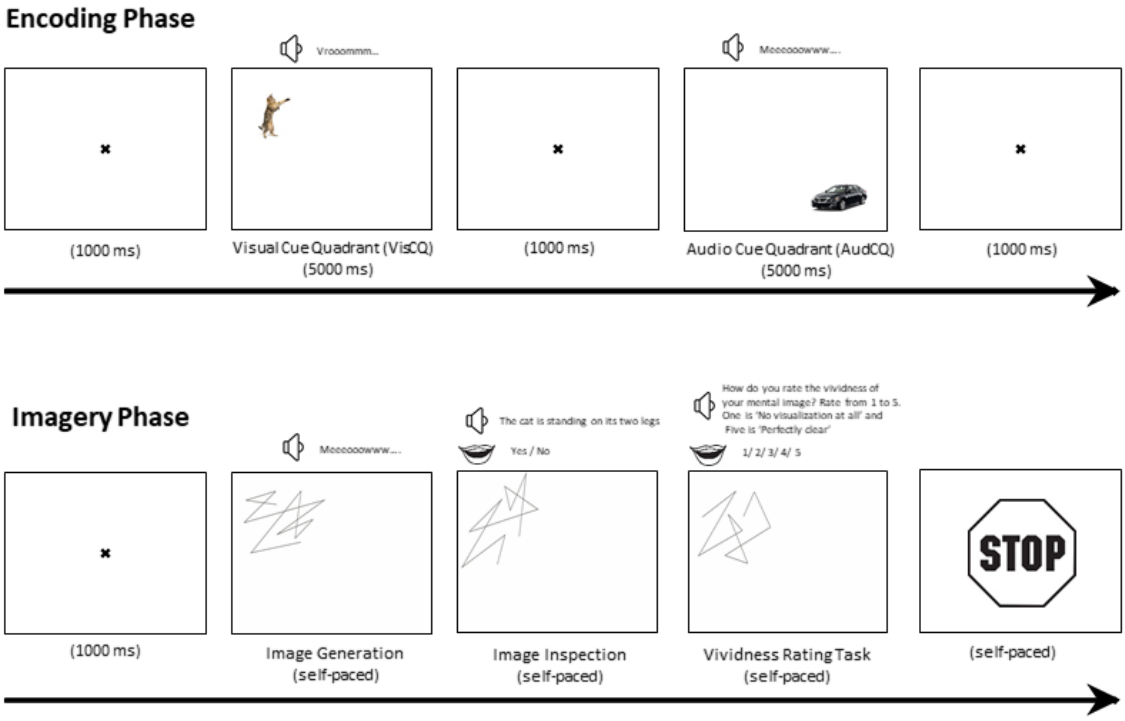


Figure 12 Illustration of the trial sequence in both experimental phases. *Encoding phase*: the figure shows an example of a 'cat-car' pairs, with the image-sound manipulation. *Imagery phase*: after the fixation screen, the participants were presented with a white blank screen together with the sound of a cat. They were required to form a mental image of a cat and maintain that mental image until they saw a 'Stop' screen. If in the 'cat' trials they attended more to the upper-left quadrant, this meant that they had a memory of the spatial orientation corresponding to the image of the cat, even though the sound did not match. However, if they attended more to the bottom-right quadrant, this meant that they had a memory of the spatial orientation corresponding to the sound of the cat, even though the image did not match.

4.2.5 Statistical Analysis

We analyzed data from the Encoding phase and the Imagery phase. In the Encoding phase, we analyzed participants' eye movements from the onset of the image-sound stimuli until their offset. In the Imagery phase, we analyzed participants' eye movements from the onset of the white blank screen until the participants' response.

Our dependent variable was the mean percentage of dwell time (i.e., the time spent with the eyes in the area of interest, AOI; Holmqvist et al., 2011). Depending on the type of stimuli encountered during the previous Encoding phase, quadrants were named as (i) Visual Cue Quadrant (hereafter, VisCQ), if the image had been shown in that quadrant; (ii) Audio Cue Quadrant (hereafter, AudCQ), if the image paired with the corresponding sound had been shown in that quadrant; and (iii) No Cue Quadrant 1 and (iv) No Cue Quadrant 2 (hereafter, NoCQ1 and NoCQ2), if no image had been shown in that quadrant (1 and 2 were attributed following the 'clock coded' method by Richardson and Spivey (2000), starting from the VisCQ).

For each participant and trial, we calculated the mean percentage of dwell time in each quadrant. This was analyzed with mixed ANOVA, with Task (Task 1, Task 2, Task 3) and Cue Quadrant (VisualCQ, AudioCQ, NoCQ1, NoCQ2) as within-subjects factors and Group (intra-categorical: Group A, extra-categorical: Group B) as between-subjects factor. By including Cue Quadrant as factor, we could assess whether dwell time during imagery varied across quadrants, depending on where the corresponding visual or auditory stimuli had been previously presented. If dwell time were longer in the VisualCQ, for instance, this would suggest that participants were re-enacting previous gaze patterns, relying more on visual information. By including

Task as factor, we could assess whether dwell time also varied across imagery tasks. In the Image Inspection task, for instance, participants also need to retrieve specific information (e.g., the tail of the cat is long, the nose of the cat is black), and longer dwell time may thus be needed (Martarelli & Mast, 2013). By including Group as a factor, we could assess the influence of distinct semantic manipulations on dwell time during spatial imagery tasks. To assess whether dwell time in each quadrant also differed from chance (i.e., 25%), we also used one-sample *t*-tests. When the assumption of sphericity was violated ($p < .05$), we used the Huynh-Feldt correction to adjust the degrees of freedom. The effect size was reported as Partial Eta Squared and Cohen's *d*. We used Tukey corrections to adjust for multiple comparisons. The alpha level for statistical tests was set at .05.

4.3 Results

4.3.1 Perceptual Encoding phase: Percentage of dwell time during stimulus presentation

To ensure that stimuli were properly encoded, in the Encoding phase we compared dwell time in the quadrants containing the stimulus, to mean dwell time in the other three quadrants.The *t*-test showed that participants spent significantly more time in the quadrant where the image stimuli were located ($M = 95.92\%$, $SD = 2.66$), as compared to the other quadrants ($M = 4.08\%$, $SD = 2.66$; $t(29) = 94.418$, $p < .001$, $d = 17.24$). Hence, the analysis confirmed that stimuli were properly encoded.

4.3.2 Imagery phase: Percentage of dwell time

In the Imagery phase, we separately analyzed correct and incorrect trials (i.e., trials in which participants provided correct/incorrect responses in the Image Inspection task). Participants mostly gave accurate responses (72.19%), which are analyzed below. Analyses of the inaccurate response (27.81%) are reported in a separate section.

The ANOVA analyses showed a non-significant interaction between Task × Cue Quadrant × Group, (F(3.83, 107.20) = 1.09, p > .05 ., η^2_p = .04). We therefore tested all the 2-way interactions, and found that they were all non-significant (Task × Cue Quadrant: F(3.83, 107.20) = 2.04, p > .05, η^2_p = .07; Cue Quadrant × Group: F(1.98, 55.37) = 0.27, p > .05, η^2_p = .01; Task × Group: F(1, 28) = .00, p > .05, η^2_p = .00). Also the main effect of Group (F(1, 28) = .00, p > .05, η^2_p = .00) and Task (F(1, 28) = .00, p > .05, η^2_p = .00) were not significant. The analyses only showed a significant main effect of Cue Quadrant on dwell time (F(1.98, 55.37) = 7.41, p < .005, η^2_p = .21). In particular, dwell time was significantly higher in VisualCQ (34.01%, *SD* 1.65) than in the other Cue Quadrants (AudioCQ: 21.64%, *SD* 0.64, p = .001; NoCQ1: 20.98%, *SD* 1.26, p < .001; NoCQ2: 23.38%, *SD* 1.53, p = .006). Mean percentages of dwell time across tasks and quadrants are illustrated in Figure 13.

Mean dwell time was significantly higher than chance in the VisualCQ of Tasks 1, 2 and 3 (Task 1: t(29) = 4.20, p < .001, d = 0.77; Task 2: t(29) = 3.31, p < .005, d = 0.60; Task 3: t(29) = 2.59, p < .05, d = 0.47). Mean dwell time in the AudioCQ was significantly below chance in all three tasks (Task 1: t(29) = 2.64, p < .05, d = 0.48; Task 2: t(29) = 3.38, p < .005, d = 0.62; Task 3: t(29) = 3.52, p < .005, d = 0.64). Mean dwell time in the NoCQ1 was at chance level in Task 1 (t(29) = 1.83, p

> .05, d = 0.33), and below chance in Tasks 2 and 3 (Task 2: $t(29) = 2.16$, $p < .05$, d = 0.40); Task 3: $t(29) = 2.82$, $p < .01$, $d = 0.51$). Finally, mean dwell time in NoCQ2 was at chance levels in all three tasks (Task 1: $t(29) = 1.60$, $p > .05$, $d = 0.29$; Task 2: $t(29) = 1.12$, $p > .05$, $d = 0.20$; Task 3: $t(29) = 0.06$, $p > .05$, $d = 0.01$).

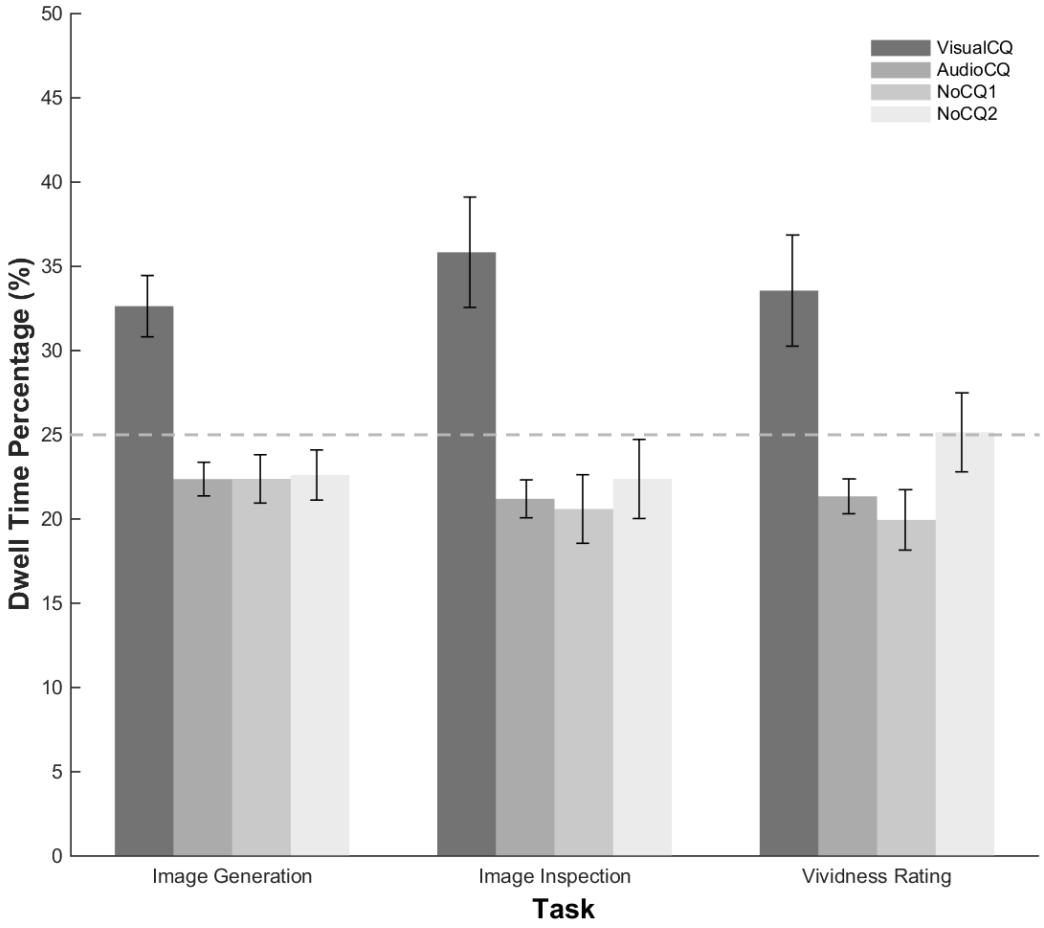


Figure 13 Graph showing the mean percentage of dwell time in each quadrant during the Image Generation task, the Image Inspection task and the Vividness Rating task in the Imagery phase. Error bars represent the standard errors of the mean. The dotted grey line represents chance levels.

4.3.3 Analyses of incorrect trials

When only analyzing incorrect trials (i.e., when participants gave inaccurate responses during the Image Inspection task), we found a significant main effect of Cue Quadrant on dwell time ($F(3, 84) = 4.12$, $p < .01$, $\eta^2_p = .13$). Post-hoc comparisons only revealed a significant difference between dwell time in VisualCQ (30.34%, *SD* 1.14) and NoCQ1 (19.24%, *SD* 1.75; $p < .01$).

One-sample *t*-tests further revealed that dwell time was significantly higher than chance in VisualCQ for the three tasks (Task 1: 29.17%, *SD* 11.40, $t(29) = 2.01$, $p = .05$, $d = 0.37$; Task 2: 30.39%, *SD* 14.17, $t(29) = 2.09$, $p < .05$, $d = 0.38$; Task 3: 31.44% *SD* 15.57, $t(29) = 2.27$, $p < .05$, $d = 0.41$). Moreover, dwell time was significantly below chance in NoCQ1 in Tasks 2 and 3 (Task 2: 18.03% *SD* 14.24, $t(29) = 2.68$, $p < .05$, $d = 0.49$; Task 3: 18.73%, *SD* 14.89, $t(29) = 2.31$, $p < .05$, $d = 0.42$).

4.4 Discussion

In this study we monitored looking patterns, to understand how visual and auditory stimuli are mentally reconstructed during mental imagery. Our results showed that when participants were instructed to generate on an empty screen the mental image corresponding to a certain sound, participants looked longer in those areas in which that particular image had been previously seen (Visual Cue Quadrant; VisCQ). In contrast, they did not look longer in those areas in which they had seen another image associated to the same sound. Dwell time was not affected by either task nor semantic manipulations.

In the Encoding phase, participants spent significantly more time in the quadrant where the visual stimuli were located, as compared to the other quadrants, confirming that the stimuli were properly encoded and our results reliable.

In the Imagery phase, participants mostly gave accurate responses. During mental imagery, they looked significantly longer in those areas in which the image had been previously seen (i.e., dwell time was longer in VisualCQ than in the other quadrants, and it was significantly higher than chance only in VisualCQ). These data suggest that during mental imagery participants were re-visiting and re-enacting the gaze pattern that they had shown when initially perceiving the image, looking longer where they had been previously looking. These results are in line with existing literature (e.g., Martarelli et al., 2017; Martarelli & Mast, 2011, 2013; Wantz et al, 2015), and provide further support to the hypothesis that eye movements during mental imagery play a functional role (Bochynska & Laeng, 2015; Laeng et al., 2014; Laeng & Teodorescu, 2002), with the re-enactment of looking patterns facilitating information retrieval during imagery. Laeng et al. (2014), for instance, found that the higher the resemblance of scanpath patterns during perception and imagery, the higher the accuracy of memory retrieval.

Moreover, our results are also in line with the modality appropriateness hypothesis (Welch & Warren, 1980). In particular, our results showed that participants relied more on previous visual (rather than auditory) information, when re-enacting gaze patterns during mental imagery. In line with our predictions, participants showed longer dwell time where visual (rather than auditory) stimuli had been previously shown, since visual stimuli are dominant in the processing of spatial information (e.g., Schneider et al., 2008; Talsma et al., 2010; Thelen et al., 2015). This may be because

visual inputs are mapped topologically onto the retina, and may be more important in spatial localization tasks. Therefore, these results suggest that visual information is more important than auditory information, in case of inter-sensory discrepancy during spatial localization tasks.

Very similar results were found when analyzing inaccurate responses, although the effect was unsurprisingly less strong (e.g., dwell time in the areas in which they had previously seen the image, VisualCQ, was not higher than in the areas in which they had previously seen another image associated to the same sound, AudioCQ).

These results were true for all tasks, and regardless of the semantic manipulations implemented (i.e., whether incongruencies were extra- or intra-categorical). Firstly, in contrast to our predictions, dwell time did not vary across imagery tasks and areas. In particular, dwell time in VisualCQ in the Image Inspection task was as long as in the other tasks, although participants were required to retrieve more specific information (e.g., the tail of the cat is long, the nose of the cat is black). Secondly, semantic manipulations of our multimodal stimuli also had no effect on dwell time. In particular, extra-categorical stimuli (e.g., an animal image paired with an object sound) are more incongruent than intra-categorical stimuli (e.g., an animal image paired with another animal sound), and may thus be even harder to process (e.g., Vogler & Titchener, 2011; Yuval-Greenberg & Deouell, 2009). However, dwell time in VisualCQ was not higher for intra-categorical stimuli, in contrast with our predictions.

4.5 Conclusion

Overall, our study confirmed previous studies showing a dominant role of the visual modality in spatial imagery tasks (as compared to the auditory modality), and suggesting a functional role of eye movements during mental imagery. These results are especially important, considering that we used different types of stimuli, as compared to those used in previous studies on spatial imagery activity. Further experimental investigations on modality effect and imagery should be undertaken in the future, in order to better explore the extent to which auditory inputs influence spatially-related memory retrieval.

Chapter 5

General Discussion

The goal of this dissertation was to investigate how multimodal information is processed, recognized and retrieved. In particular, I studied (i) how attention is allocated toward congruent and incongruent multimodal stimuli (Chapter 2), (ii) how multimodal stimuli are recognized in recognition memory tasks (Chapter 3), and (iii) how they are retrieved during spatial imagery activity (Chapter 4). Processing multimodal stimuli is a complex phenomenon (Mayer & Anderson, 1991; Mayer & Sims, 1994), which importantly differs from the processing of unimodal stimuli (e.g., Dunifon et al., 2016; Thompson & Paivio, 1994; Goolkasian & Foos, 2005). Therefore, studying how multimodal stimuli are processed, recognized and retrieved is essential, given that we live in a multisensory world and we continuously receive sensory inputs from multiple modalities.

The aim of the first study (Chapter 2) was to understand how context and modality interplay during the allocation of visual attention. Our study showed that visual attention is allocated differently, depending on the context, congruency and modality of the stimuli used. These factors interplay in a complex way, and incongruities between visual and auditory stimuli result in different attention allocation between target objects and contexts. Participants, for instance, allocated more attention to the target object in case of congruent auditory stimuli or incongruent visual stimuli, while the opposite was true when allocating attention to the context. These results are important, because they show that different stimuli are processed in

different ways, and the effect of modality should be better taken into account when studying how attention is allocated between target stimuli and context.

The aim of the second study (Chapter 3) was to investigate how the modality of the stimuli used affects the ability to identify familiar information (i.e., recognition memory; Kafkas & Montaldi , 2015; Võ et al., 2008). Our results showed significant differences in recognition memory, depending on the novelty and modality of the stimuli used. In particular, novel auditory stimuli were the hardest to be processed, in contrast to our predictions. These findings are important, because they show that old stimuli are not always harder to process than novel ones, as it instead happens in the visual modality (e.g., Kafkas & Montaldi, 2017; Rugg & Curran, 2007; Võ et al., 2008). In the auditory modality, indeed, novel stimuli may be harder to process, possibly because they are especially relevant for humans and largely monopolize their attention.

Finally, the aim of the third study (Chapter 4) was to investigate how visual and auditory stimuli are mentally reconstructed in the absence of corresponding sensory stimulations (i.e., mental imagery; Lacey, 2013). Our study showed that when participants had to generate mental images corresponding to a certain sound, they looked longer in those areas in which that particular image had been previously seen, but not in the areas where they had seen another image associated to the same sound. These findings confirm the dominant role of the visual modality in spatial imagery tasks (as compared to the auditory modality), in line with the modality appropriateness hypothesis (Welch & Warren, 1980), and further suggest that eye movements during mental imagery have a functional role.

Our results largely confirm previous findings, showing a general prevalence of the visual modality in spatial tasks, at least during mental imagery tasks. As discussed by Dunifon et al. (2016), various studies in the last 40 years have shown the dominance of the visual modality in a variety of different tasks (e.g., Colavita, 1974). However, such dominance effect is not rigid, as the modality dominance is also affected by the quality, characteristics and contextual circumstances of the sensory stimulation (Talsma et al., 2010; Yuval-Greenberg & Deouell, 2009).

Our results also open up to new lines of research, by showing for instance that, in the auditory modality, novel stimuli may be harder to process than old ones, at least in recognition memory tasks. These results are important, because they challenge the common view that old stimuli are generally harder to process (e.g., Kafkas & Montaldi, 2015, 2017, b; Montefinese, Ambrosini, Fairfield, & Mammarella, 2013; Rugg & Curran, 2007; Võ et al., 2008). Clearly, these findings will need to be validated by more experiments in the future.

Similarly, our results showed a complex interaction of congruency and modality on the way visual attention was allocated between target objects and contexts. When allocating attention between congruent and incongruent multimodal stimuli, for instance, more attention was allocated to the target objects not only in case of incongruent visual stimuli, but also in case of congruent auditory stimuli. This complex interplay of context, modality and congruency is something that should be better taken into account in future studies.

Overall, the findings support the relevance of modality appropriateness hypothesis (Welch & Warren, 1980), and the flexibility of the dominance effect (with the characteristics and contextual circumstances of the stimuli importantly affecting performance; see (Talsma et al., 2010; Yuval-Greenberg & Deouell, 2009). Moreover, they challenge the idea that old stimuli are generally harder to process, and confirm the importance of eye-tracking data to study implicit cognitive processing.

Another important point that should be highlighted is that from all the three studies that have been conducted, the most obvious finding to emerge is that each sensory modality manipulations were found to elicits different cognitive and oculomotor response. The oculomotor response made for each visual and auditory manipulations seems to follow an asymmetrical pattern, in which each modality was found to demonstrate different response that is not parallel with each other. This finding is interesting as it shows the uniqueness of each sensory modalities and how responsive the modality is towards different kinds of manipulations.

These findings add to the growing body of multimodal research on the bidirectional influences in information processing between visual and auditory modalities. The novelty that have been introduced in this dissertation is that the experimental design for all the studies conducted have incorporates the use of multimodal stimulus that are not very common in the discussion of those particular studies. In addition, the original contributions to knowledge is all the studies conducted have utilized different kinds of stimulus that is distinct from stimulus of previous studies. All studies conducted used multimodal stimulus so the modality effect that each modality manipulation brings can be examine.

Finally, it is important to note that these experiments were limited by the absence of other behavioural and physiological measures. In particular, the interpretation of the data and the conclusions made were largely based on eye-tracking measures. Despite these limitations, the experiments conducted certainly add to our understanding of how multimodal stimuli are processed, recognized and retrieved.

In the future, more studies should follow this approach, complementing eye-tracking data with multimodal stimulation and, ideally, with neurophysiological data and other behavioural measures (e.g., reaction time). While the use of eye-tracking data will provide objective measures of how humans process, recognize and retrieve stimuli, the use of multimodal stimuli will ensure an ecologically more valid set-up, allowing researchers to draw stronger conclusions on these complex processes.

References

Andersen, R. A., Snyder, L. H., Bradley, D. C., & Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience, 20*(1), 303-330.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4*, 829.

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science, 13*(3), 99-102.

Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance, 19*(2), 250-267.

Ballas, J. A., & Mullins, T. (1991). Effects of context on the identification of everyday sounds. *Human Performance, 4*(3), 199-219.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*(8), 617-629.

Beck, D. M., & Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Research, 49*(10), 1154-1165.

Blajenkova, O., Kozhevnikov, M., & Motes, M. A. (2006). Object-spatial imagery: A new self-report imagery questionnaire. *Applied Cognitive Psychology, 20*(2), 239-263.

Bochynska, A., & Laeng, B. (2015). Tracking down the path of memory: Eye scanpaths facilitate retrieval of visuospatial information. *Cognitive Processing, 16*(1), 159-163.

Borji, A., Sihite, D.N., & Itti, L. (2014). What/where to look next? Modeling top-down visual attention in complex interactive environments. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44*(5), 523-538.

Botvinick, M. M., Braver, T. S., Barch, D. M., & Carter, C. S. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624-652.

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

Brisson, J., Mainville, M., Mailloux, D., Beaulieu, C., Serres, J., & Sirois, S. (2013). Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers. *Behavior Research Methods, 45*(4), 1322-1331.

Broadbent, D. E. (1957). A mechanical model for human attention and immediate memory. *Psychological Review, 64*(3), 205-215.

Brocher, A., & Graf, T. (2016). Pupil old/new effects reflect stimulus encoding and decoding in short-term memory. *Psychophysiology, 53*(12), 1823-1835.

Brocher, A., & Graf, T. (2017). Response: Commentary: Pupil old/new effects reflect stimulus encoding and decoding in short-term memory. *Frontiers in Psychology, 8*, 539.

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.

Chen, Y., Nguyen, T., V., Kankanhalli, M., Yuan, J., Yan, S., & Wang, M. (2014). Audio matters in visual attention. *IEEE Transactions on Circuits and Systems for Video Technology, 24*(11), 1992-2003.

Chen, Y., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition, 114*(3), 389-404.

Chen, Y., & Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance, 37*(5), 1554-1568.

Chun, M. M. (2000). Contextual cueing of visual attention. In (Vol. 4, pp. 170-178): Elsevier Ltd.

Coco, M. L., Malcolm, G. L., & Keller, F. (2014). The interplay of bottom-up and top-down mechanisms in visual guidance during object naming. *Quarterly Journal of Experimental Psychology, 67*(6), 1096-1120.

Cohen, R. A. (2013). The neuropsychology of attention. In (2 ed.). Boston, MA: Springer.

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics, 16*(2), 409-412.

Colonius, H., & Diederich, A. (2006). The race model inequality: interpreting a geometric measure of the amount of violation. *Psychological Review, 113*(1), 148.

Curran, T., & Friedman, W. J. (2004). ERP old/new effects at different retention intervals in recency discrimination tasks. *Cognitive Brain Research, 18*(2), 107-120.

Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition, 35*(3), 393-401.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15*(8), 559-564.

Delogu, F., Raffone, A., & Belardinelli, M. O. (2009). Semantic encoding in working memory: Is there a (multi) modality effect? *Memory, 17*(6), 655-663.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*(1), 193-222.

Diaconescu, A. O., Alain, C., & McIntosh, A. R. (2011). The co-occurrence of multisensory facilitation and cross-modal conflict in the human brain. *Journal of Neurophysiology, 106*(6), 2896-2909.

Duchowski, A. T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers, 34*(4), 455-470.

Dunifon, C. M., Rivera, S., & Robinson, C. W. (2016). Auditory stimuli automatically grab attention: Evidence from eye tracking and attentional manipulations. *Journal of Experimental Psychology: Human Perception and Performance, 42*(12), 1947.

Fiedler, A. (2013). Redundancy gain for semantic features. *Psychonomic Bulletin & Review, 20*(3), 474-480.

Fort, A., Delpuech, C., Pernier, J., & Giard, M.-H. (2002). Dynamics of Cortico-subcortical Cross-modal Operations Involved in Audio-visual Object Detection in Humans. *Cerebral Cortex, 12*(10), 1031-1039.

Freides, D. (1974). Human information processing and sensory modality: Cross-modal functions, information complexity, memory, and deficit. *Psychological Bulletin, 81*(5), 284-310.

Giard, M. H., & Peronnet, F. (1999). Auditory-Visual Integration during Multimodal Object Recognition in Humans: A Behavioral and Electrophysiological Study. *Journal of Cognitive Neuroscience, 11*(5), 473-490.

Gomes, C. A., Montaldi, D., & Mayes, A. (2015). The pupil as an indicator of unconscious memory: Introducing the pupil priming effect. *Psychophysiology, 52*(6), 754-769.

Goolkasian, P., & Foos, P. W. (2005). Bimodal format effects in working memory. *The American Journal of Psychology, 118*(1), 61-78.

Greene, A. J., Easton, R. D., & LaShell, L. S. R. (2001). Visual-auditory events: Cross-modal perceptual priming and recognition memory. *Consciousness and Cognition, 10*(3), 425-435.

Hales, J. B., & Brewer, J. B. (2011). The timing of associative memory formation: Frontal lobe and anterior medial temporal lobe activity at associative binding predicts memory. *Journal of Neurophysiology, 105*(4), 1454.

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

Henderson, J. M. (1992). Object identification in context: The visual processing of natural scenes. *Canadian Journal of Psychology/ Revue Canadienne de Psychologie, 46*(3), 319-341.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics, 57*(6), 787-795.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*: OUP Oxford.

Irwin, D. E. (2004). Fixation location and fixation duration as indices of cognitive processing. In J. M. H. F. Ferreira (Ed.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 105-133). New York, NY, US: Psychology Press.

Johansson, R., Holsanova, J., & Homqvist, K. (2011). *The dispersion of eye movements during visual imagery is related to individual differences in spatial imagery ability.* Paper presented at the Proceedings of the Cognitive Science Society.

Kafkas, A., & Montaldi, D. (2015). The pupillary response discriminates between subjective and objective familiarity and novelty. *Psychophysiology, 52*(10), 1305-1316.

Kafkas, A., & Montaldi, D. (2017). Commentary: Pupil old/new effects reflect stimulus encoding and decoding in short-term memory. *Frontiers in Psychology, 8*, 277.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154*(3756), 1583-1585.

Kahneman, D., & Peavler, W. S. (1969). Incentive effects and pupillary changes in association learning. *Journal of Experimental Psychology, 79*(2, pt.1), 312-318.

Kloosterman, N. A., Meindertsma, T., van Loon, A. M., Lamme, V. A. F., Bonneh, Y. S., & Donner, T. H. (2015). Pupil size tracks perceptual content and surprise. *European Journal of Neuroscience, 41*(8), 1068-1078.

Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica, 134*(3), 372-384.

Küper, K., Groh-Bordin, C., Zimmer, H. D., & Ecker, U. K. H. (2012). Electrophysiological correlates of exemplar-specific processes in implicit and explicit memory. *Cognitive, Affective, & Behavioral Neuroscience, 12*(1), 52-64.

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

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

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

Lai, M.-L., Tsai, M.-J., Yang, F.-Y., Hsu, C.-Y., Liu, T.-C., Lee, S. W.-Y., . . . Tsai, C.-C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review, 10*, 90-115.

LaPointe, M. R. P., Lupianez, J., & Milliken, B. (2013). Context congruency effects in change detection: Opposing effects on detection and identification. *Visual Cognition, 21*(1), 99-122.

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research, 158*(4), 405-414.

Lewandowski, L. J., & Kobus, D. A. (1993). The effects of redundancy in bimodal word processing. *Human Performance, 6*(3), 229-239.

Lewkowicz, D. J. (1988a). Sensory dominance in infants: I. Six-month-old infants' response to auditory-visual compounds. *Developmental Psychology, 24*(2), 155.

Lewkowicz, D. J. (1988b). Sensory dominance in infants: II. Ten-month-old infants' response to auditory-visual compounds. *Developmental Psychology, 24*(2), 172.

Lukas, S. (2009). *Cross-modal selective attention in switching stimulus modalities*. Hamburg: Kovač.

Lukas, S., Philipp, A. M., & Koch, I. (2010). Switching attention between modalities: Further evidence for visual dominance. *Psychological Research, 74*(3), 255-267.

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

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

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

Mast, F. W., & Kosslyn, S. M. (2002). Eye movements during visual mental imagery. *Trends in Cognitive Sciences, 6*(7), 271-272.

Mayer, R. E., & Anderson, R. B. (1991). Animations need narrations: An experimental test of a dual-coding hypothesis. *Journal of Educational Psychology, 83*(4), 484.

Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology, 86*(3), 389.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology, 14*(2), 247-279.

Min, X., Zhai, G., Gao, Z., Hu, C., & Yang, X. (2014, 18-20 Sept. 2014). *Sound influences visual attention discriminately in videos.* Paper presented at the 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX).

Mishra, R. K. (2015). The many shades of attention. In *Interaction between attention and language systems in humans* (pp. 21-55): Springer India.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex, 14*(4), 452-465.

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013). The "subjective" pupil old/new effect: Is the truth plain to see? *International Journal of Psychophysiology, 89*(1), 48-56.

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology, 94*(1), 156.

Mudrik, L., Deouell, L. Y., & Lamy, D. (2011). Scene congruency biases Binocular Rivalry. *Consciousness and Cognition, 20*(3), 756-767.

Naber, M., Frässle, S., Rutishauser, U., & Einhäuser, W. (2013). Pupil size signals novelty and predicts later retrieval success for declarative memories of natural scenes. *Journal of Vision, 13*(2), 11-11.

Nava, E., & Pavani, F. (2013). Changes in sensory dominance during childhood: Converging evidence from the Colavita effect and the sound-induced flash illusion. *Child Development, 84*(2), 604-616.

Özcan, E., & van Egmond, R. (2009). The effect of visual context on the identification of ambiguous environmental sounds. *Acta Psychologica, 131*(2), 110-119.

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology, 83*(1), 56-64.

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review, 83*(2), 157.

Ralph, B. C. W., Seli, P., Cheng, V. O. Y., Solman, G. J. F., & Smilek, D. (2014). Running the figure to the ground: Figure-ground segmentation during visual search. *Vision Research, 97*, 65-73.

Ramdane-Cherif, Z., & NaÏt-AliNait-Ali, A. (2008). An adaptive algorithm for eye-gaze-tracking-device calibration. *IEEE Transactions on Instrumentation and Measurement, 57*(4), 716-723.

Reinwein, J. (2012). Does the modality effect exist? And if so, which modality effect? *Journal of Psycholinguistic Research, 41*(1), 1-32.

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

Robinson, C. W., & Sloutsky, V. M. (2004). Auditory Dominance and Its Change in the Course of Development. *Child Development, 75*(5), 1387-1401.

Röder, B., & Büchel, C. (2009). Multisensory interactions within and outside the focus of visual spatial attention (Commentary on Fairhall & Macaluso). *European Journal of Neuroscience, 29*(6), 1245-1246.

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences, 11*(6), 251-257.

Ryan, J. D., Hannula, D. E., & Cohen, N. J. (2007). The obligatory effects of memory on eye movements. *Memory, 15*(5), 508-525.

SanMiguel, I., Linden, D., & Escera, C. (2010). Attention capture by novel sounds: Distraction versus facilitation. *European Journal of Cognitive Psychology, 22*(4), 481-515.

Schneider, T. R., Engel, A. K., & Debener, S. (2008). Multisensory identification of natural objects in a two-way crossmodal priming paradigm. *Experimental Psychology, 55*(2), 121-132.

SensoMotoric Instruments. (2009). *iView X System Manual (Version 2.4). Teltow, Germany.* In.

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research, 14*(1), 147-152.

Sinnett, S., Soto-Faraco, S., & Spence, C. (2008). The co-occurrence of multisensory competition and facilitation. *Acta Psychologica, 128*(1), 153-161.

Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: The Colavita effect revisited. *Perception & Psychophysics, 69*(5), 673-686.

Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(6), 679-692.

Sloutsky, V. M., & Napolitano, A. C. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child Development, 74*(3), 822-833.

Sloutsky, V. M., & Robinson, C. W. (2008). The role of words and sounds in infants' visual processing: From overshadowing to attentional tuning. *Cognitive Science, 32*(2), 342-365.

Stubblefield, A., Jacobs, L., Kim, Y., & Goolkasian, P. (2013). Colavita dominance effect revisited: The effect of semantic congruity. *Attention, Perception, & Psychophysics, 75*(8), 1827-1839.

Suied, C., Bonneel, N., & Viaud-Delmon, I. (2009). Integration of auditory and visual information in the recognition of realistic objects. *Experimental Brain Research, 194*(1), 91-102.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences, 13*(9), 403-409.

Talsma, D. (2015). Predictive coding and multisensory integration: An attentional account of the multisensory mind. *Frontiers in Integrative Neuroscience, 9*, 19.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences, 14*(9), 400-410.

Thelen, A., Talsma, D., & Murray, M. M. (2015). Single-trial multisensory memories affect later auditory and visual object discrimination. *Cognition, 138*, 148-160.

Thompson, V. A., & Paivio, A. (1994). Memory for pictures and sounds: Independence of auditory and visual codes. *Canadian Journal of Experimental Psychology/ Revue Canadienne de Psychologie Expérimentale, 48*(3), 380.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520.

Tsilionis, E., & Vatakis, A. (2016). Multisensory binding: Is the contribution of synchrony and semantic congruency obligatory? *Current Opinion in Behavioral Sciences, 8*, 7-13.

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *The Quarterly Journal of Experimental Psychology, 59*(11), 1931-1949.

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition, 17*(1), 159-170.

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review, 25*(6), 2005-2015.

Vandierendonck, A. (2016). Modality independence of order coding in working memory: Evidence from cross-modal order interference at recall. *The Quarterly Journal of Experimental Psychology, 69*(1), 161-179.

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

Võ, M. L. H., & Henderson, J. M. (2011). Object-scene inconsistencies do not capture gaze: Evidence from the flash-preview moving-window paradigm. *Attention, Perception, & Psychophysics, 73*(6), 1742-1753.

Võ, M. L. H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., & Hutzler, F. (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology, 45*(1), 130-140.

Vogler, J. N., & Titchener, K. (2011). Cross-modal conflicts in object recognition: Determining the influence of object category. *Experimental Brain Research, 214*(4), 597-605.

Voss, J. L., & Paller, K. A. (2008). Brain substrates of implicit and explicit memory: The importance of concurrently acquired neural signals of both memory types. *Neuropsychologia, 46*(13), 3021-3029.

Wantz, A. L., Martarelli, C. S., & Mast, F. W. (2015). When looking back to nothing goes back to nothing. *Cognitive Processing, 17*(1), 105-114.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*(3), 638-667.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods, 42*(3), 671-684.

Wixted, J. T. (2009). Remember/Know judgments in cognitive neuroscience: An illustration of the underrepresented point of view. *Learning & Memory, 16*(7), 406.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*(3), 441-517.

Yun, K., Peng, Y., Samaras, D., & Zelinsky, G. J. (2013). Exploring the role of gaze behavior and object detection in scene understanding. *Frontiers in Psychology, 4*, 917.

Yuval-Greenberg, S., & Deouell, L. Y. (2009). The dog's meow: Asymmetrical interaction in cross-modal object recognition. *Experimental Brain Research, 193*(4), 603.

Zelinsky, G. J. (2013). Understanding scene understanding. *Frontiers in Psychology, 4*.

List of Appendix

Appendix A	Letter from the University of Bern institutional ethics review board
Appendix B	List of visual and auditory stimuli used in Experiment 1
Appendix C	List of visual (geometrical shapes) and auditory (sounds) stimuli used in Experiment 2
Appendix D	List of visual and auditory stimuli used in Experiment 3
Appendix E	Debriefing for Experiment 1
Appendix F	Debriefing for Experiment 2
Appendix G	Debriefing for Experiment 3
Appendix H	Inform consent form
Appendix I	Participants profile form
Appendix J	Research poster

Appendix A: Letter from the University of Bern institutional ethics review board

UNIVERSITÄT BERN

An
Frau Dr. Corinna Martarelli
Institut für Psychologie
Universität Bern
Fabrikstrasse 8
3012 Bern

Philosophisch-humanwissenschaftliche Fakultät

Ethikkommission

Bern, 17. Dezember 2015

Ihr Ethikantrag zum Projekt „How the eyes view audio-visual irregularities"

Nr. 2015-12-000004

Sehr geehrte Frau Doktor Martarelli

die Ethikkommission hat Ihren o.g. Antrag geprüft und die darin beschriebene Studie als ethisch unbedenklich eingestuft.

Mit freundlichen Grüssen

Prof. Dr. Thomas Berger
Präsident der Ethikkommission

Prof. Dr. Thomas Berger
Präsident der Ethikkommission
Fabrikstr. 8
CH-3012 Bern

Tel. +41 31 631 34 07
Fax +41 31 631 82 12
thomas.berger@ptp.unibe.ch

Appendix B: List of visual and auditory stimuli used in Experiment 1

	Image of:-	Sound of:-	Situation of:-
Congruent Context, Congruent Sound (CC)	**Animate** 1. Cat 2. Dog 3. Chicken 4. Rooster 5. Duck 6. Goose 7. Cow 8. Goat 9. Tiger 10. Elephant 11. Horse 12. Bird 13. Bee 14. Monkey 15. Frog **Inanimate** 1. Car 2. Bicycle 3. Helicopter 4. Ship 5. Fire Truck 6. Aeroplane 7. Piano 8. Guitar (string) 9. Drum (percussion) 10. Blender 11. Telephone 12. Microwave 13. Vacuum 14. Washing machine 15. Lawn mower	1. Cat 2. Dog 3. Chicken 4. Rooster 5. Duck 6. Goose 7. Cow 8. Goat 9. Tiger 10. Elephant 11. Horse 12. Bird 13. Bee 14. Monkey 15. Frog 1. Car 2. Bicycle 3. Helicopter 4. Ship 5. Fire Truck 6. Aeroplane 7. Piano 8. Guitar (string) 9. Drum (percussion) 10. Blender 11. Telephone 12. Microwave 13. Vacuum 14. Washing machine 15. Lawn Mower	1. Living room 2. House yard 3. Farm 4. Countryside 5. Pond 6. Farm/ Yard 7. Pasture field 8. Farm 9. Forest 10. African grassland 11. Race track 12. Tree & Nest 13. Bee hive 14. Tree & Forest 15. Pond 1. Highway 2. Road 3. Helipad on rooftop 4. Ocean 5. Fire department 6. Airport 7. Living hall 8. Music studio 9. Stage 10. Kitchen 11. Office desk 12. Kitchen 13. Living room 14. Laundry room 15. Garden

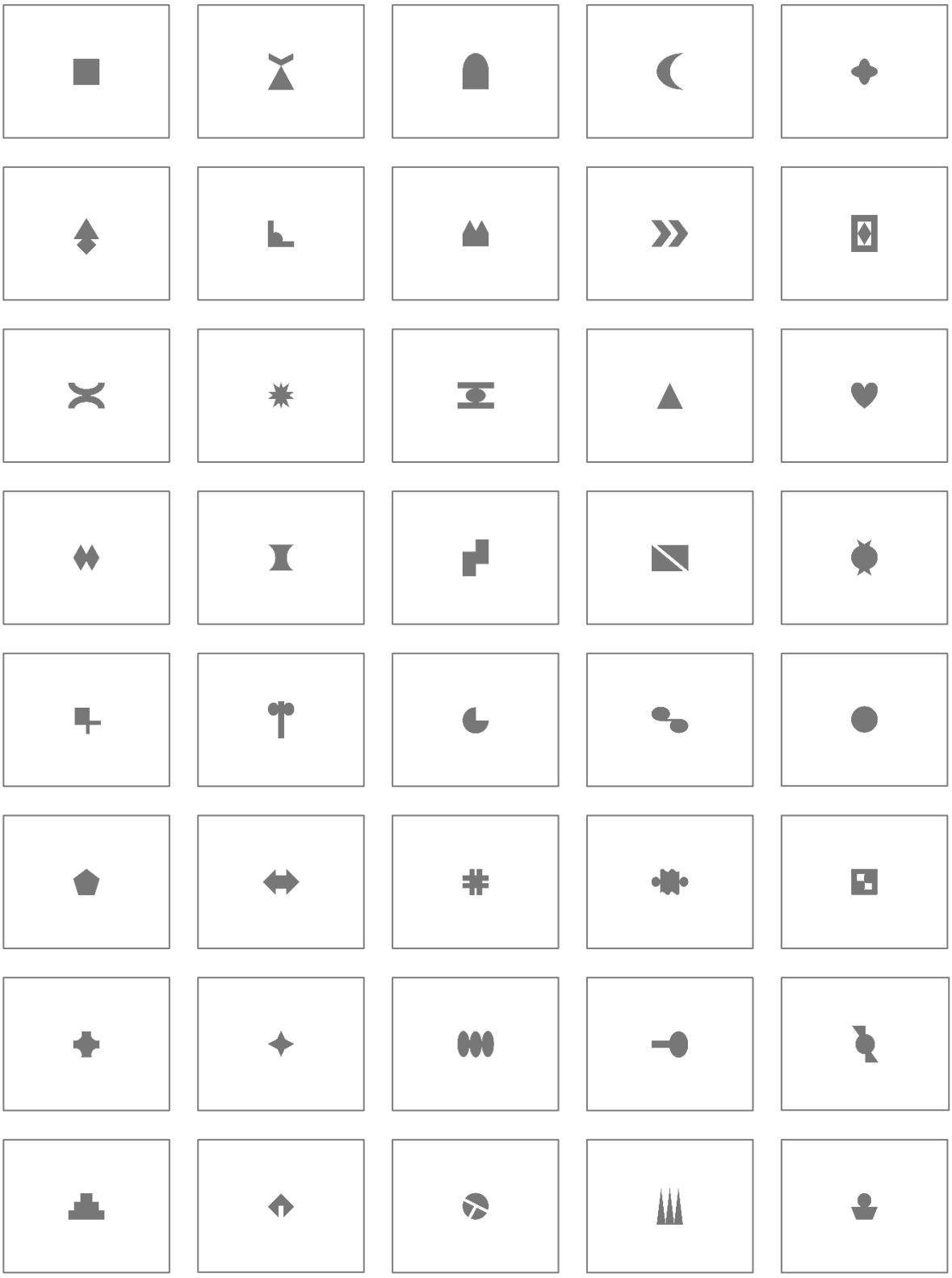
	Image of:-	Sound of:-	Situation of:-
Congruent Context, Incongruent Sound (CI)	**Animate** 1. Cat 2. Dog 3. Chicken 4. Rooster 5. Duck 6. Goose 7. Cow 8. Goat 9. Tiger 10. Elephant 11. Horse 12. Bird 13. Bee 14. Monkey 15. Frog **Inanimate** 1. Car 2. Bicycle 3. Helicopter 4. Ship 5. Fire Truck 6. Aeroplane 7. Piano 8. Guitar (string) 9. Drum (percussion) 10. Blender 11. Telephone 12. Microwave 13. Vacuum 14. Washing machine 15. Lawn mower	1. Chicken 2. Duck 3. Cat 4. Cow 5. Sheep 6. Goat 7. Horse 8. Dog 9. Monkey 10. Tiger 11. Frog 12. Bee 13. Bird 14. Rooster 15. Elephant 1. Ship 2. Aeroplane 3. Bicycle 4. Ambulance 5. Helicopter 6. Train 7. Drum 8. Trumpet 9. Piano 10. Microwave 11. Vacuum 12. Alarm clock 13. Broom 14. Telephone 15. Bell	1. Living room 2. House yard 3. Farm 4. Countryside 5. Pond 6. Farm/ Yard 7. Pasture field 8. Farm 9. Forest 10. African grassland 11. Race track 12. Tree & Nest 13. Bee hive 14. Tree & Forest 15. Pond 1. Highway 2. Road 3. Helipad on rooftop 4. Ocean 5. Fire department 6. Airport 7. Living hall 8. Music studio 9. Stage 10. Kitchen 11. Office desk 12. Kitchen 13. Living room 14. Laundry room 15. Garden

Cont.: **List of visual and auditory stimuli used in Experiment 1**

	Image of:-	Sound of:-	Situation of:-
Incongruent Context, Congruent Sound (IC)	**Animate** 1. Cat 2. Dog 3. Chicken 4. Rooster 5. Duck 6. Goose 7. Cow 8. Goat 9. Tiger 10. Elephant 11. Horse 12. Bird 13. Bee 14. Monkey 15. Frog **Inanimate** 1. Car 2. Bicycle 3. Helicopter 4. Ship 5. Fire Truck 6. Aeroplane 7. Piano 8. Guitar (string) 9. Drum (percussion) 10. Blender 11. Telephone 12. Microwave 13. Vacuum 14. Washing machine 15. Lawn mower	1. Cat 2. Dog 3. Chicken 4. Rooster 5. Duck 6. Goose 7. Cow 8. Goat 9. Tiger 10. Elephant 11. Horse 12. Bird 13. Bee 14. Monkey 15. Frog 1. Car 2. Bicycle 3. Helicopter 4. Ship 5. Fire Truck 6. Aeroplane 7. Piano 8. Guitar (string) 9. Drum (percussion) 10. Blender 11. Telephone 12. Microwave 13. Vacuum 14. Washing machine 15. Lawn Mower	1. Factory 2. Roller coaster 3. Living room 4. Water surface 5. Restaurant 6. Bowling arena 7. Market 8. Classroom 9. Supermarket 10. Train station 11. Petrol station 12. Office 13. Under the sea 14. Post office 15. Café 1. Railway track 2. Kitchen 3. Under the sea 4. Mountain 5. Swimming pool water surface 6. Hotel 7. Tennis court 8. Library 9. Hospital 10. Bedroom 11. Vegetable stand 12. Bathroom 13. House lawn 14. Playground 15. Living Room

	Image of:-	Sound of:-	Situation of:-
Incongruent Context, Incongruent Sound (II)	**Animate** 1. Cat 2. Dog 3. Chicken 4. Rooster 5. Duck 6. Goose 7. Cow 8. Goat 9. Tiger 10. Elephant 11. Horse 12. Bird 13. Bee 14. Monkey 15. Frog **Inanimate** 1. Car 2. Bicycle 3. Helicopter 4. Ship 5. Fire Truck 6. Aeroplane 7. Piano 8. Guitar (string) 9. Drum (percussion) 10. Blender 11. Telephone 12. Microwave 13. Vacuum 14. Washing machine 15. Lawn mower	1. Chicken 2. Duck 3. Cat 4. Cow 5. Sheep 6. Goat 7. Horse 8. Dog 9. Monkey 10. Tiger 11. Frog 12. Bee 13. Bird 14. Rooster 15. Elephant 1. Ship 2. Aeroplane 3. Bicycle 4. Ambulance 5. Helicopter 6. Train 7. Drum 8. Trumpet 9. Piano 10. Microwave 11. Vacuum 12. Alarm clock 13. Broom 14. Telephone 15. Bell	1. Factory 2. Roller coaster 3. Living room 4. Water surface 5. Restaurant 6. Bowling arena 7. Market 8. Classroom 9. Supermarket 10. Train station 11. Petrol station 12. Office 13. Under the sea 14. Post office 15. Café 1. Railway track 2. Kitchen 3. Under the sea 4. Mountain 5. Swimming pool water surface 6. Hotel 7. Tennis court 8. Library 9. Hospital 10. Bedroom 11. Vegetable stand 12. Bathroom 13. House lawn 14. Playground 15. Living room

Appendix C: List of visual (geometrical shapes) and auditory (sounds) stimuli used in Experiment 2



Cont.: **List of auditory stimuli used in Experiment 2**

Animate sounds:

- bee.mp3
- bird.mp3
- cat.mp3
- chicken.mp3
- cow.mp3
- crickets.mp3
- crow.mp3
- dog2.mp3
- dolphin.mp3
- duck.mp3
- elephant.mp3
- frog.mp3
- goat.mp3
- goose.mp3
- horse.mp3
- monkey.mp3
- rooster.mp3
- tiger.mp3
- turkey.mp3
- wolf.mp3

Inanimate sounds:

- aeroplane.mp3
- bagpipe.mp3
- bicycle.mp3
- blender.mp3
- car.mp3
- drum.mp3
- guitar.mp3
- hair dryer.mp3
- helicopter.mp3
- lawn mower.mp3
- microwave.mp3
- piano.mp3
- printer.mp3
- ship.mp3
- tambourine.mp3
- telephone.mp3
- train.mp3
- trumpet_ic.mp3
- vacuum.mp3
- washing machine.mp3

Appendix D: List of visual and auditory stimuli used in Experiment 3

Group A: Extra-Categorial

No.	Image	Sound
1	Cat Car	Car Cat
2	Lion Hand bell	Hand bell Lion
3	Monkey Guitar	Guitar Monkey
4	Donkey Police car	Police car siren Donkey
5	Duck Scissors	Scissors Duck
6	Sheep Vacuum	Vacuum Sheep
7	Cow Washing machine	Washing machine Cow
8	Wolf Bagpipes	Bagpipes Wolf
9	Flies Train	Train Flies
10	Snake Airplane	Airplane Snake
11	Owl Bicycle	Bicycle Owl
12	Rooster Tambourine	Tambourine Rooster
13	Eagle Telephone	Telephone Eagle
14	Mouse Flute	Flute Mouse
15	Bear Drum	Drum Bear
16	Crickets Electric drill	Electric drill Crickets

Group B: Intra-Categorial

No.	Image	Sound
1	Dog Chicken	Chicken Dog
2	Tiger Horse	Horse Tiger
3	Pig Frog	Frog Pig
4	Elephant Mosquito	Mosquito Elephant
5	Camel Dolphin	Dolphin Camel
6	Goose Crocodile	Crocodile Goose
7	Canary Goat	Goat Canary
8	Turkey Bee	Bee Turkey
9	Ship Fire truck	Fire truck siren Ship
10	Piano Trumpet	Trumpet Piano
11	Helicopter Motorcycle	Motorcycle Helicopter
12	Violin Bongos	Bongos Violin
13	Whistle Hair dryer	Hair dryer Whistle
14	Lawn mower Hammer	Hammer Lawn mower
15	Alarm clock Blender	Blender Alarm clock
16	Printer Microwave	Microwave Printer

Appendix E: Debriefing for Experiment 1

Debriefing
How the eyes view audio-visual irregularities

Dear participants,

Thank you for your participation in this eye-tracking experiment.

This study investigates how cross-modal incongruity affects human visual cognition. Previous studies have shown that humans are more interested in events that violate their expectation. It has been illustrated that incongruent visual stimuli (pictures that contain incongruent objects) elicit longer fixation durations compared to congruent visual stimuli (Underwood & Foulsham, 2006; Underwood, Templeman, Lamming, & Foulsham, 2008; Ralph, Seli, Cheng, Solman, & Smilek, 2014).

We are interested in the effects of visual but also auditory incongruity (mismatch between visual and auditory information) on eye behavior and pupil size. For example, we expect a smaller pupil size with both visual and auditory incongruent trials. Indeed, the pupil is a reliable physiological marker of novelty (Kahneman, 1973). More generally, this is an explorative study that will help us to better understand which modality is more sensitive to incongruity. Will the effect of incongruity be more important in the visual or auditory modality?

References

Kahneman, D. (1973). *Attention and effort*. Engelwood Cliffs, NJ: Prentice Hall.

Ralph, B. C. W., Seli, P., Cheng, V. O. Y., Solman, G. J. F., & Smilek, D. (2014). Running the figure to the ground: Figure-ground segmentation during visual search. *Vision Research, 97*, 65-73.

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *The Quarterly Journal of Experimental Psychology, 59*(11), 1931-1949.

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition, 17*(1), 159-170.

If you have any questions or are interested in the results,
please contact Hafidah Umar, hafidah.binti-umar@psy.unibe.ch

MANY THANKS…!!!

Appendix F: Debriefing for Experiment 2

Debriefing
Name of Study: Don't be confused..!!!

Dear participants,

Thank you for your participation in this eye-tracking experiment.

You just completed two experimental studies which are both related to the topic on cross-modal conflicts. Below is the brief description on those studies.

Experiment 2 (The shapes and sounds)

This study investigates how the congruity and incongruity of audio-visual stimulation affects the pupil response. The congruent stimuli refers to the stimuli which have been presented during the learning phase (the old stimuli) and the incongruent stimuli is the stimuli which have not been presented before (the new stimuli).

The term that is frequently used in the literature to represent this line of study is 'pupil old-new effect'. The main idea behind this notion is that our pupil can discriminates between the old stimuli and the new stimuli. In which the old stimuli will elicits larger pupil dilation than the new stimuli (Kafkas & Montaldi, 2015; Montefinese, Ambrosini, Fairfield, & Mammarella, 2013). Extensive research has shown that our pupil react to a strength of memory signal, and pupillometry is a good techniques to explore the underlying mechanism of recognition memory (Otero, Samantha, Brendan, & Samuel, 2011).

Despite many previous studies shows that pupil diameter increased when people viewed old (congruent) items compared to new (incongruent) items, very little is known about how pupil old-new effect relates to the audio-visual stimulation. Is pupil old-new effect also applies to the combination of audio-visual stimuli?

References

Kafkas, A., & Montaldi, D. (2015). The pupillary response discriminates between subjective and objective familiarity and novelty. *Psychophysiology, 52*(10), 1305-1316.

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013). The "subjective" pupil old/new effect: Is the truth plain to see? *International Journal of Psychophysiology, 89*(1), 48-56.

Otero, S. C., Samantha, C. O., Brendan, S. W., & Samuel, B. H. (2011). Pupil size changes during recognition memory. *Psychophysiology, 48*(10), 1346-1353.

If you have any questions or are interested in the results,
please contact Hafidah Umar, hafidah.binti-umar@psy.unibe.ch

MANY THANKS…!!!

Appendix G: Debriefing for Experiment 3

Debriefing
Draw Your Mind Out

Dear participants,

You just completed one experimental session.

Thank you for your participation in this eye-tracking experiment.

Below is the brief description on the study.

This study is related to the topic of cross-modal visuo-spatial imagery. We are interested to investigate further how did the audio-visual incongruity affects the eye gazing behavior during mental imagery activity. This study is based on many studies on visuo-spatial imagery which suggest that the eye movement during perception and image generation phase shares the same pattern (Martarelli, Chiquet, Laeng, & Mast, 2016; Bochynska & Laeng, 2015). Perception refers to our ability to see, hear, or become aware of something through the senses. While imagery can easily be defined as visualization. It occurs whenever a person has a conscious sensory experience, but in reality there is no physical or real stimulation (Lacey & Lawson, 2013).

In discussing about the visual and space, one important idea is the functional theory of image generation. According to this theory, the visual system re-enacts the same oculomotor behavior that occurred at encoding and this oculomotor behavior assists the construction of the mental image. Or in other words, it suggest that the eye movement pattern during the perception and imagery shares the same pattern (Richardson & Spivey, 2000).

Generally, this study is explorative in nature in which it allows us to gain insight on which modality have more influence on visual imagery activity. Are people more incline to visualize in visual field where the image match or the sound match.

References

Bochynska, A., & Laeng, B. (2015). Tracking down the path of memory: eye scanpaths facilitate retrieval of visuospatial information. *Cognitive Processing, 16*(1), 159-163.

Lacey, S., & Lawson, R. (2013). *Multisensory imagery*: Springer New York.

Martarelli, C. S., Chiquet, S., Laeng, B., & Mast, F. W. (2016). Using space to represent categories: insights from gaze position. *Psychological Research*, 1-9.

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

If you have any questions or are interested in the results, please contact Hafidah.

E-mail: hafidah.binti-umar@psy.unibe.ch (OR) hafidah1006@gmail.com

MANY THANKS…!!!

Appendix H: Inform consent form

Einverständniserklärug

UNIVERSITÄT BERN

Bitte lesen Sie dieses Formular sorgfältig durch.

Bitte fragen Sie den/die Untersucher/in oder Ihre Kontaktperson, wenn Sie etwas nicht verstehen oder etwas wissen möchten.

ProbandIn (Vor- und Nachname) : _____

UntersucherIn (Vor- und Nachname) : _____

i. Ich nehme freiwillig an dieser Studie teil.

ii. Ich kann meine Mitarbeit an dieser Studie jederzeit, ohne Angabe von Gründen, abbrechen.

iii. Ich bin über den Aufbau und die Zielsetzung, über die zu erwartenden Wirkungen, über mögliche Vor- und Nachteile sowie über eventuelle Risiken der Studie unterrichtet worden.

iv. Ich bin damit einverstanden, dass alle aufgezeichneten Daten unter Wahrung meiner Anonymität aufbewahrt und ausgewertet werden und für wissenschaftliche (und Ausbildungs-) Zwecke verwendet werden.

v. Ich nehme zur Kenntnis, dass ich innerhalb der nächsten 6 Monate verlangen kann, dass meine persönlichen Daten permanent gelöscht werden.

Unterschrift Proband/in : _____

Ort, Datum : _____

Unterschrift Untersucher/in : _____

ID:

Appendix I: Participants profile form

ID:

Profile Form

Instructions: Please fill in the required information and circle the CORRECT response

First Name	:	_____
Last Name	:	_____
Gender	:	Male ☐ Female ☐
Are you a psychology student?	:	Yes / No
Year of Study	:	1st year / 2nd year / 3rd year / 4th year
Date of Birth	:	_____
Age	:	_____
Nationality	:	_____
First language	:	_____
Second language	:	_____
Handedness	:	_____
Visual acuity	:	Normal / Corrected
If corrected, in what form	:	Glasses / Contact lenses / LASIK
Right now I am wearing	:	Glasses / Contact lenses
Color blindness	:	Yes. I am color blind / No. I am not color blind

Appendix J: **Research poster: Presented at SGS-CCLM Summer School at Weggis, Switzerland (26th June, 2016)**

How the Eyes View Audio-Visual Irregularities

Hafidah Umar[1,2], Benjamin Steinweg[1], Trix Cacchione[1], Fred Mast[1], Corinna Martarelli[1]

1) Institute of Psychology, University of Bern, Switzerland and Center for Cognition, Learning and Memory, University of Bern, Switzerland
2) Center for Neuroscience Services & Research (P3Neuro), Universiti Sains Malaysia

Research Background & Objectives

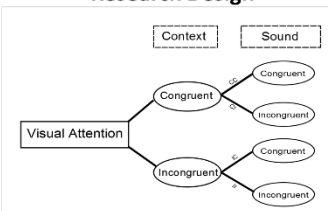
Our complex environment provides us with information from different sensory modalities that are processed and integrated based on our knowledge of the real world. To date, studies on *multisensory integration* have been showing that congruity between modalities facilitates performance (for example in a detection task), whereas incongruity interferes with performance (e.g. Chen et al., 2014). It has also been illustrated that humans are more attracted by events that violate expectations (Ralph, Seli, Cheng, Solman, & Smilek, 2014).

For example, it has been shown that incongruent stimuli elicit longer fixations (Underwood & Foulsham, 2006; Underwood, Templeman, Lamming, & Foulsham, 2008). Here, we aim to investigate not only the effects of visual but also auditory incongruity via an *eye-tracking* approach. Which modality (visual or auditory) is more sensitive to congruity-incongruity effects? More generally, how do our eyes behave when they encounter a visual and/or auditory incongruent situation?

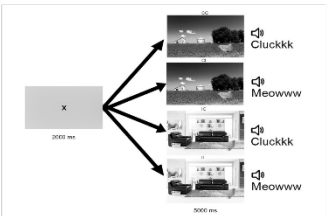
Research Methods

- 34 participants (mean age = 23.18)
- 2x2 within-subject experimental design
- 120 audio-visual stimuli
- Luminance of the images was controlled with the SHINE toolbox in MATLAB.
- We also controlled the position of the target object (same across conditions)
- Rating (manipulation check)
- Qualitative /explicit question

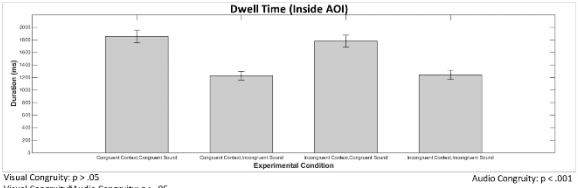
Research Design



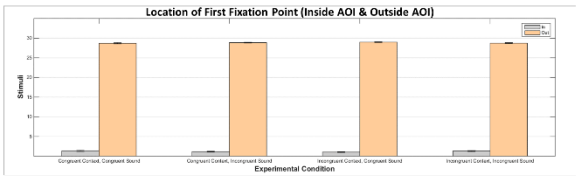
Trial Procedure

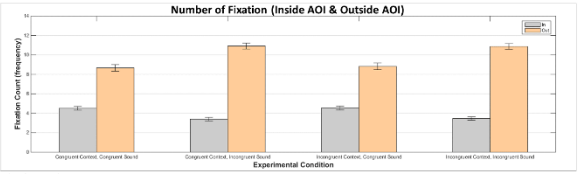


Results

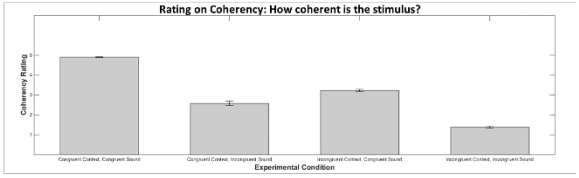


Visual Congruity: p > .05
Visual Congruity*Audio Congruity: p > .05
Audio Congruity: p < .001

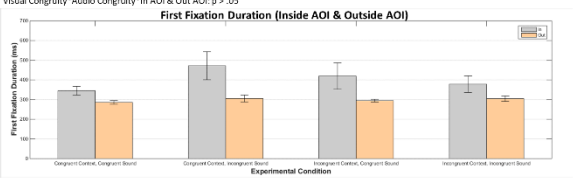




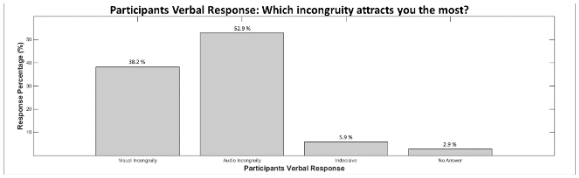
Visual Congruity: p > .05
Visual Congruity*Audio Congruity: p > .05
Visual Congruity*In AOI & Out AOI: p > .05
Visual Congruity*Audio Congruity*In AOI & Out AOI: p > .05
Audio Congruity: p < .001
In AOI & Out AOI: p < .001
Audio Congruity*In AOI & Out AOI: p < .001



Visual Congruity: p < .001
Audio Congruity: p < .001
Visual Congruity*Audio Congruity: p < .001



Visual Congruity: p > .05
Audio Congruity: p > .05
Visual Congruity*Audio Congruity: p > .05
Visual Congruity*In AOI & Out AOI: p > .05
Audio Congruity*In AOI & Out AOI: p > .05
Visual Congruity*Audio Congruity*In AOI & Out AOI: p > .05
In AOI & Out AOI: p < .05



Conclusion

The more implicit measure of eye behavior showed that participants spent more time on the target-object when the target-object is congruent with the sound regardless of scene context. This result is congruent with Chen et al. (2014). When the target-object is incongruent with the sound the number of fixations outside the AOI (not on the target-object) is higher compared to when the target-object is congruent with the sound. This suggests a more active searching behavior when there is a mismatch between visual and auditory information.

The analysis of verbal responses revealed that most of the participants agreed that the audio incongruity is more attractive compared to visual incongruity. We found a correspondence between the more implicit measure of eye behavior and the explicit responses.

Our data provide evidence that the audio modality plays an important role in directing people visual attention.

References:-

Chen, Y., Nguyen, T.V., Kankanhalli, M., Yuan, J., Yan, S., & Wang, M. (2014). Audio matters in visual attention. *IEEE Transactions on Circuits and Systems for Video Technology, 24* (11): 1992-2003.

Ralph, B. C. W., Seli, P., Cheng, V.O.Y., Solman, G.J.F., & Smilek, D. (2014). Running the figure to the ground: Figure-ground segmentation during visual search. *Vision Research*, 97: 65-73.

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *The Quarterly Journal of Experimental Psychology, 59* (11): 1931-1949.

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition, 17*: 159-170.

e-mail: hafidah.binti-umar@psy.unibe.ch