

Multisensory Processing: The Formation of Object Categories

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy to Trinity College Dublin, the University of Dublin

Martina A. Seveso

B. Sc., M. Sc.

19/01/2026

Under the supervision of Professor Fiona N. Newell

School of Psychology and Institute of Neuroscience



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Declaration

I, Martina Seveso, declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. I confirm that the work has been completed in full compliance with relevant university policies including the Academic, Integrity policy, the Trinity Policy on Good Research Practice, and the guidance on AI and GenAI in Teaching, Learning, Assessment and Research.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

I have read and understood the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at <http://www.tcd.ie/calendar>. I have also read and understood the guide, and completed the 'Ready Steady Write' Tutorial on avoiding plagiarism, located at <https://libguides.tcd.ie/academic-integrity/ready-steady-write>.

Summary

Every day we categorise myriad objects using information from multiple sensory modalities. For instance, when we encounter a cat, we perceive it through its visual properties (e.g., shape, size, colour), the sounds it makes (e.g., meowing), and its texture (e.g., by petting its fur). Indeed, real-world object perception is inherently multisensory (Shams & Seitz, 2008), involving concurrent visual, auditory, and tactile inputs. Although there is abundant evidence for multisensory interactions in the brain (Ghazanfar & Schroeder, 2006; Alais et al., 2010), most investigations into object categorisation to date have focused mainly on unimodal sensory processing. Therefore, the processes underlying the integration of multisensory cues in categorisation and generalisation remain poorly understood (Newell et al., 2023).

The aim of this thesis was to investigate the role of multisensory perception in the formation of object categories, and to examine whether multisensory information leads to more robust categorisation and more efficient generalisation. To this end, we explored the contribution of visual, auditory, and tactile information to the formation of multisensory perceptual categories.

In Chapter 2, we examined whether prior audio-visual exposure influenced subsequent visual categorical perception of familiar objects. The results provided no evidence that auditory object cues facilitated category representation of familiar objects. However, participants successfully categorised stimuli on the basis of visual shape alone, even in the absence of categorical perception. Chapter 3 investigated whether motion and sound cues influenced the perceived shape of unfamiliar objects, as assessed through judgements of inter-object similarity. The results suggested that visual motion cues systematically affected the perceived similarity and classification of unfamiliar objects, whereas auditory motion cues did not. In Chapter 4, we examined whether visual and tactile motion cues influenced the categorisation of unfamiliar objects, and whether the formation of these multisensory categories subsequently benefited

generalisation to novel exemplars. The findings provided compelling evidence that both visual and tactile motion cues supported more efficient categorisation and generalisation of unfamiliar objects. Finally, in Chapter 5, we investigated the resolution of visuo-haptic conflict in a categorisation task conducted in virtual reality. The results revealed haptic dominance in shape-based categorisation when both visual and haptic shape information were available.

Taken together, these findings indicate differential effectiveness of distinct sensory cues (i.e., spatial versus temporal information). While visual shape cues benefited from simultaneous access to multiple spatial features (i.e., visual and haptic inputs), auditory cues, which rely primarily on temporal dynamics, did not provide sufficiently diagnostic information to support categorisation. In Chapter 6, the findings of this thesis are discussed in relation to their theoretical implications, the gaps in the literature addressed by this experimental work, and potential directions for future research.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Professor Fiona Newell, for her guidance, constructive feedback and consistent support throughout the course of my PhD. Thank you for showing me the ‘arts of the carpenter’, I hope I have built a solid chair. I am also thankful to my thesis committee members, Professor Lorina Naci and Professor Anna Borghi, for their time, careful review, and the insightful discussion. Secondly, I am also grateful to the researchers I have been honoured to collaborate with in a formative and inspiring way: Dr Rebecca Hirst, Nottingham Trent University, Nottingham, for showing me the secrets and beauties of PsychoPy; Dr Alan O’Dowd, Trinity College Dublin, for our theoretical and statistical conversations and his patience in teaching model approaches; and Prof. Ivan Camponogara, Zayed University, Abu Dhabi, for his expertise and statistical precision which enriched this research. This research would not have been possible without the support of Science Foundation Ireland (SFI, grant number 19/FFP/6812), funding training and conferences attendance, and the Department of Psychology at Trinity College Dublin, providing guidance and resources.

On a personal note, I am grateful to all the people who supported me in countless and meaningful ways during my PhD, who became my home away from home. Eimear, whom I shared this experience with, thank you for your constant support and encouragement; Al, for our walks-talks and shared passion for art; Kate, for making me laugh daily – Christmas paper will always have a special meaning to us; Nicole, for sharing a love of *good* food; Nina, for reminding me to shake it off and xyz. Thanks to Jack, who tried to teach me Irish – *Go n-éirí an bóthar leat*; Isabella, for your positive and beautiful soul; Jade, for yoga, mats and early breakfasts; Harvey, for your endless style. Finally, thanks to Dr Fabiana Lombardi for our enduring friendship - the picture in white coats is now reality; Ana and George, for showing me how to kart, be kind and there for

each other; Alessio and Ana, for many Italian aperitivos. My gratitude also goes to all the other friends who supported me remotely during this journey.

Finally, I am deeply grateful to my family: my mother Cristina, who always believed in me even during the most difficult times; my father Armando ever by my side, who taught me to work hard and always aim higher; Olives, who brought a significant amount of joy into my life; and to my partner Jean-Christophe Michel Florent, who endlessly supported me with laughter, challenges, and care – you pushed me to believe in my ideas and your support has been a constant source of motivation and perseverance.

Publications

The following are the manuscripts that have been submitted for publication through this Ph.D.

Publications arising from this thesis (to date):

Newell, F. N., McKenna, E., **Seveso, M. A.**, Devine, I., Alahmad, F., Hirst, R. J., & O'Dowd, A. (2023). Multisensory perception constrains the formation of object categories: A review of evidence from sensory-driven and predictive processes on categorical decisions. *Philosophical Transactions of the Royal Society B*, 378(1886), 20220342. ([Chapter 1](#))

Seveso, M.A., Hirst, R.J., O'Dowd, A., Camponagaro I. & Newell, F.N. (2026) Visual and tactile motion cues enhance the categorisation of novel object shapes. *Experimental Brain Research*. ([Chapter 4](#))

Seveso, M.A., Kidney E., & Newell, F.N. (2025, *Under review*) Evidence for a Haptic Over Vision Bias in the Categorisation of Novel Objects using Virtual Reality. *Transactions on Applied Perception*. ([Chapter 5](#))

Table of contents

Declaration

Summary

Acknowledgements

Publications

List of Figures

List of Tables

1	Introduction to Multisensory Objects Categorisation.....	2
1.1	Overview	2
1.2	The Multisensory Brain	4
1.2.1	Theories of Multisensory Integration	4
1.2.2	Behavioural evidence of Multisensory Integration	7
1.2.3	Neurophysiological and Neuroimaging Evidence of Multisensory Integration.....	9
1.2.4	The Early versus Late Multisensory Integration Debate.....	11
1.3	Categorisation of Unisensory Objects	12
1.3.1	Visual Object Categorisation	12
1.3.1.1	Neural Models of Visual Object Perception: Recognition and Categorisation	16
1.3.2	Haptic Object Categorisation.....	17
1.3.2.1	Neural models of haptic objects: recognition and categorisation	20
1.3.3	Auditory object Categorisation.....	21
1.3.3.1	Neural models of auditory objects: recognition and categorisation	23
1.4	Categorisation of Multisensory Object	24
1.4.1	Visuo-auditory interactions in object perception and categorisation.....	24
1.4.1.1	Visuo-auditory interactions in the perception of novel objects	26
1.4.2	Visuo-haptic interactions in object perception and categorisation	28
1.4.2.1	Visuo-haptic interactions in object perception of novel objects	29
1.4.3	Audio-haptic interactions in object perception and categorisation.....	30
1.5	Outline and Scope of this Thesis	32

2	Audio-Visual Categorical Perception of Familiar Objects.....	35
2.1	Abstract.....	35
2.2	Introduction.....	36
2.3	Outline of the Experiments	40
2.4	General Stimuli	41
2.5	Experiment 1A	44
2.5.1	Method	45
2.5.2	Result	48
2.5.3	Discussion	52
2.6	Experiment 1B	53
2.6.1	Method	55
2.6.2	Result	64
2.6.3	Discussion	75
2.7	Experiment 1C	78
2.7.1	Method	78
2.7.2	Result	81
2.7.3	Discussion	85
2.8	General Discussion	86
2.9	Conclusion	95
3	The Way an Object Moves, but not its Sound, Influences its Perceived Similarity to Other Object Shape.....	98
3.1	Abstract.....	98
3.2	Introduction.....	99
3.3	Outline of the Experiments	102
3.4	General Stimuli	103
3.5	Experiment 2A	107
3.5.1	Method	108
3.5.2	Result	113
3.5.3	Discussion	119
3.6	Experiment 2B	120
3.6.1	Method	121
3.6.2	Result	122
3.6.3	Discussion	129

3.7 Experiment 3.....	132
3.7.1 Method.....	133
3.7.2 Result	135
3.7.3 Discussion.....	139
3.8 General Discussion	140
3.9 Conclusions.....	143
4 Visual and Tactile Motion Cues Enhance the Categorisation of Novel Object Shapes: Evidence from Remote Testing Using Mobile Devices	146
4.1 Abstract.....	146
4.2 Introduction.....	147
4.3 Outline of the Experiments	149
4.4 Experiment 4.....	150
4.4.1 Method.....	151
4.4.2 Result	159
4.4.3 Discussion.....	161
4.5 Experiment 5.....	163
4.5.1 Method.....	163
4.5.2 Result	165
4.5.3 Discussion.....	167
4.6 Experiment 6.....	168
4.6.1 Method.....	168
4.6.2 Result	169
4.6.3 Discussion.....	170
4.7 Exploratory Analyses of the Category Boundary.....	170
4.7.1 Results.....	172
4.7.2 Discussion.....	175
4.8 General Discussion	176
4.9 Conclusions.....	180
5 A Crossmodal Shape Conflict Results in a Haptic Over Vision Bias in the Categorisation of Novel 3D Objects: a Virtual Reality Study.....	183
5.1 Abstract.....	183
5.2 Introduction.....	184
5.3 Outline of the Experiment	189

5.4	Experiment 7.....	190
5.4.1	Method	190
5.4.2	Result	199
5.4.3	Discussion	205
5.5	Conclusion	209
6	General Discussion	212
6.1	Introduction.....	212
6.2	Summary of Main Findings.....	213
6.3	Implications of the Findings.....	214
6.3.1	The visual object shape, but not its sound, support the categorisation of familiar objects 214	
6.3.2	The role of visual object shape and audio-visual object motion cues on inter-object perceived similarity of unfamiliar objects	216
6.3.3	Beyond Shape: the role of visuo-tactile motion cues in categorising multisensory unfamiliar objects	218
6.3.4	Haptic dominance in shape conflict for categorisation of unfamiliar objects	220
6.3.5	Theoretical implications: toward a continuum approach for multisensory object representations.....	221
6.3.5.1	Task demand specificity and information type	224
6.4	Novel Contribution of this Thesis	226
6.5	Limitations of the Current Work and Future Directions	227
6.6	Conclusion	229
7	References	230
8	Appendices	284
8.1	Appendix A - Ethics Approval letters	284
8.2	Appendix B – Experimental Supplementary Materials	286

List of Figures

Figure 1 An illustration of the Categorical Perception phenomenon. _____	37
Figure 2 An illustration of the familiar objects adopted in Experiment A, 1B and 1C, grouped by basic level category and the object shapes continua. _____	43
Figure 3 Example of an experimental trial in Experiment 1A. _____	47
Figure 4 Overall performance for the sound-object exemplar matching task in Experiment 1A. _____	49
Figure 5 Overall performance for the sound-object exemplar matching task in Experiment 1A. _____	51
Figure 6 Overview of the experimental paradigm in Experiment 1B. _____	57
Figure 7 An illustration of a trial sequence from the audio-visual congruent learning condition in Experiment 1B. _____	60
Figure 8 Mean categorisation and discrimination responses by object pairing (A), and by learning condition (B) in Experiment 1B. _____	65
Figure 9 Probability of a category 'A' response across learning conditions for each object pair in Experiment 1B. _____	66
Figure 10 Discrimination accuracy across learning conditions for each object pair in Experiment 1B. _____	68
Figure 11 The plot shows the estimated mean discrimination accuracy predicted from the linear mixed-effects model for object pairings in Experiment 1B. _____	70
Figure 12 The plot shows the estimated slope values predicted from the linear mixed-effects model for object pairings across learning conditions in Experiment 1B. _____	71
Figure 13 Mean categorisation and discrimination reaction times by object pairing (A), and by learning condition (B) in Experiment 1B. _____	72
Figure 14 Plots showing estimated mean discrimination reaction times predicted from the linear mixed-effects model for object pairings and position In Experiment 1B. _____	75
Figure 15 The plot shows normalised similarity ratings responses for Experiment 1C. _____	84

Figure 16 An illustration of the 'shape space' of object stimuli used in the Experiments 2 and 3.

_____ 105

Figure 17 A schematic illustration of the timeline of a typical trial in Experiment 2 (from left to right). _____ 111

Figure 18 The plots show the main results based on similarity ratings in Experiment 2A. ____ 116

Figure 19 The plots show the two-dimensional non-metric MDS solutions in Experiment 2A. _ 118

Figure 20 The plot shows the summary of the main findings from linear mixed-effects models predicting probability of similarity ratings in Experiment 2B. _____ 125

Figure 21 The plot shows the main predictions from the linear mixed-effects models on the probability of similarity ratings for the motion condition across steps size in Experiment 2B. _ 127

Figure 22 An illustration of two trial examples from a speeded classification task where participants judge whether two objects share the same feature along a relevant dimension in Experiment 3. _____ 135

Figure 23 Summary plots performance and interference effect of reaction times in milliseconds in Experiment 3. _____ 138

Figure 24 An illustration of the 'shape space' of object stimuli used in the Experiments 4, 5 and 6. _____ 152

Figure 25 A schematic illustration of the sequence of events from stimulus presentation to the response display used in (A) the learning session and (B) the categorisation test in Experiments 4 and 5. _____ 155

Figure 26 Plots showing estimated mean categorisation accuracy predicted from the binomial mixed-effects model for the different cue conditions and for each exemplar type in Experiment 4. _____ 160

Figure 27 Estimated mean categorisation accuracy predicted from the binomial mixed-effects model for the different cue conditions and for each exemplar type in Experiment 5, in which shape similarity was a less predictive cue of category membership. _____ 166

Figure 28 Estimated mean categorisation accuracy predicted from the binomial mixed-effects model for the different cue conditions and for each exemplar type in Experiment 6. _____	170
Figure 29 Plot showing the mean categorisation responses to objects positioned closest to the category boundary in Experiment 4 (A and B) and 5 (A and B). _____	172
Figure 30 The plot shows the main effect of Cue Condition on the slope values to responses at the category boundary in Experiment 4 (A and B) and 5 (A and B). _____	175
Figure 31 An illustration of the 3D objects adopted in Experiment 7, arranged by stimuli set. _____	192
Figure 32 A schematic drawing showing the task environment in Experiment 7. _____	194
Figure 33 The plots show the raw data, and the logistic regression fits on participants' categorisation responses across Blocks in Experiment 7. _____	199
Figure 34 The plot shows the category responses to the trials in which the V-H conflict information crossed the category boundary in Experiment 7. _____	201
Figure 35 The distributions of questionnaire scores across participants in Experiment 7. _____	205

List of Tables

Table 1 Sound-object category matching task performance for object category in Experiment 1A.	49
Table 2 Sound-object exemplar matching task performance for object exemplars in Experiment 1A.	50
Table 3 Mean similarity ratings and reaction times (in milliseconds) across steps for each object category in Experiment 1C.	82
Table 4 Mean normalised similarity ratings between pairs of objects across the step sizes and motion conditions in Experiment 2A.	114
Table 5 Mean normalised similarity ratings between pairs of objects across the sound and motion conditions for each step size in Experiment 2B.	123
Table 6 Mean response times to each of the sound and motion conditions across object step sizes Experiment 2B.	128
Table 7 Mean error rates for each classification condition (baseline, correlation, filtering) and relevant dimension (shape, motion) as percentages in Experiments 3.	137
Table 8 Average proportions responses by strategy clusters in Experiment 7.	203
Table 9 Proportion of responses by strategy clusters across conflict levels in Experiment 7.	203

Chapter 1

Introduction to Multisensory Objects Categorisation

1 Introduction to Multisensory Objects Categorisation

1.1 Overview

Every day we interact with various objects, and while doing so, we primarily rely on our perceptual systems. The human brain can efficiently recognise those objects, thereby maintaining perceptual consistency, regardless of variations in sensory input. Furthermore, we are able to classify objects into categories based on previous knowledge. These abilities are fundamental for our survival. For example, every driver has experienced the relief of finding their car after a search in a parking lot full of similar cars. But one may wonder how we efficiently recognise our car among many of others in a busy parking lot. It is possible that this identification is supported by a set of distinguishing features, such as the car's colour, shape and size. These so-called '*diagnostic features*' refer to the set of distinctive features that distinguish an object as belonging to a category (Rosh & Mervis, 1975). The integration of sensory information into a holistic mental representation of 'my car' enables us to efficiently search for and recognise it. However, diagnostic features for a single representation can occur across different sensory modalities, such as the car's shape and colour, as well the sound the car makes when we use the key to unlock it. Because we inhabit such a multisensory environment, object perception needs to be a dynamic process relying on integration of sensory information from multiple modalities (Ernst & Bühlhoff, 2004). Furthermore, the integration of multisensory inputs enables not only identification, but high-level cognitive functions also support meaningful object categorisation. This process is influenced by prior knowledge and memory, providing the conceptual framework to classify an object and infer its properties, functions, and relationships to other objects (Rosch et al., 1976; Freedman & Miller, 2008; Metzger & Drewing, 2019). Thus, our brain is organised in a multisensory manner that efficiently enables the integration of information across the senses to recognise and categorise objects (Ghazanfar & Schroeder, 2006; Stein & Stanford, 2008).

This chapter explores the existing literature on how the brain is organised according to multisensory principles, how our sensory systems work together to create perception of the world, and how the brain resolves the ambiguity in this process. In doing so, we first examine how information from different sensory modalities is integrated, elucidating the principles and processes underlying multisensory integration, providing supporting behavioural and neuroimaging evidence. We then briefly elucidate how individual sensory modalities (such as, vision, haptic and audition while not including taste and olfaction) process and categorise object information, addressing experimental findings and theoretical models. In the attempt to unify the two broad literature bodies on object categorisation and multisensory perception, we then focus on behavioural and neuroimaging evidence for multisensory object categorisation, separately for familiar and unfamiliar objects, while excluding patient and animal studies. By considering a wide range of empirical and theoretical evidence, we create a deeper understanding of how our brains process and integrate sensory inputs to create coherent precepts and categories, building the foundation of this thesis.

1.2 The Multisensory Brain

Every moment of each day, we are constantly bombarded with an abundance of sensory inputs from the world around us. While traditionally studied in isolation, the sensory systems of vision, hearing, and touch are now understood to be deeply interconnected (Ghazanfar & Schroeder, 2006). This integration of information from multiple sensory modalities is known as multisensory integration (MI), a fundamental principle of neural organisation and function in the brain (Stein & Stanford, 2008). MI is considered to be adaptive (Ernst & Bulthoff, 2004), as it can result in either enhanced or decreased response compared to its individual sensory components (Stein & Meredith, 1993), and is regulated by both spatial and temporal rules and the principle of inverse effectiveness (Stein & Meredith, 1993; Holmes & Spence, 2005; Starke et al., 2020a; Stein et al., 2020; Wang et al., 2020). Further evidence of MI arises from the behavioural research, showing that multisensory stimuli lead to improved performance reflected in increased detection and faster reaction times (Lehmann & Murray, 2005; Murray et al., 2005). However, the brain faces many challenges when perceiving multisensory inputs, such as correctly integrating redundant signals as well as discarding task-irrelevant information (Newell, 2004; Woods & Newell, 2004) or determining the causal structure of inputs and their sources (Chen & Spence, 2017). Various theories have been proposed to explain the constancy of MI despite those challenges.

1.2.1 Theories of Multisensory Integration

Evidence has suggested that in specific circumstances, sensory dominance takes place. More specifically, the visual dominance account posits that visual input often exerts a stronger influence on our perception than other senses, particularly when sensory information is conflicting (Rock & Victor, 1964; Sandhu & Duson, 2012; Diaconescu et al., 2013). The visual dominance account is supported by vast amount of behavioural evidence (Posner et al., 1976; McGurk & MacDonald, 1976; Howard & Templeton, 1966; Botvinick & Cohen, 1998). A well-known example of this is the

Colavita effect, in which participants respond more to the visual input of an audio-visual stimulus, when bimodally stimulated (Colavita, 1974). Moreover, the visual dominance was also reported for object perception through visual and haptic inputs (Rock & Vicktor, 1964; Power & Graham, 1976). However, some studies reported tactile rather than visual dominance (Hershberger & Misceo, 1996; Lederman & Abbot, 1981). This contrast in findings led to the development of an alternative hypothesis to better explain the multisensory integration mechanisms, referred to as the Modality Appropriateness hypothesis (MAP; Welch & Warren, 1980).

The Modality Appropriateness hypothesis posits that each modality's influence on multisensory integration depends on the modality's appropriateness for the given task. For example, vision is chosen as a better influence on spatial localisation tasks, while audition and touch have a more precise temporal perception and thus are more 'appropriate' for temporal tasks (Welch & Warren, 1980). Evidence for the MAP account has been reported across many studies (Welch et al., 1986; Atkiss et al., 2001). However, the MAP hypothesis has been challenged by findings showing that modality dominance is not fixed but depends on the relative reliability of sensory signals. For instance, audition can dominate vision in spatial localisation task when visual information is sufficiently degraded (Alais & Burr, 2004).

Finally, Ernst (2006) proposed a unified account of these two hypotheses based on optimal integration, suggesting that the perception of objects spatial features constantly relies on multisensory fusing of all available sensory signals weighted by their relative reliability (Ernst, 2006). In other words, the multisensory signals contributing to cue integration might differ in their individual reliabilities. To produce the most unbiased estimation with the smallest degree of variance, the individual sensory information is weighted relative to its reliability according to maximum-likelihood estimation (MLE; Ernst & Banks, 2002). Evidence for this has been vastly reported (Gepshtein & Banks, 2003; Alais & Burr, 2004; Knill & Pouget, 2004; Helbing & Ernst, 2007; Fetsch et al., 2009; Angelaki et al., 2009; Burge et al., 2010).

In the past decade, the prevailing approach assumes that cue integration can be explained by the Bayesian Modelling framework (Ernst, 2006), which has been proposed as a unified approach to explain multisensory object perception and decision making. This approach reflects the flexibility of using multisensory information when needed and discard irrelevant information arising from an independent origin. Additionally, the Bayesian Causal Inference (BCI) model considers both bottom-up and top-down influences and is based on the adoption of a Bayesian Inference (BI) and the Coupling Prior (Cp; Körding et al., 2007; Colonius & Diederich, 2020; Shams & Beierholm, 2022). Indeed, the brain deals with a high amount of information from different senses and computes them optimally based on the inferred causality or information reliability from each modality. The Bayesian Cp is based on the prior probability that the co-occurring information or signals from different sensory modalities may have a common cause, or in other words, that they would belong to the same object, and thus multisensory integrated. Hence, the coupling prior refers to the participant's ability to representing ambiguity between two inputs (Helbig & Ernst, 2007). Evidence for the multisensory cue integration process following the BCI has been reported in many psychophysical experiments (Körding et al., 2007; Shams et al., 2005; Wozny et al., 2010). For instance, when individuals need to address trial-varying reliability and discrepancy of cross-modal (audio-visual) information, they integrate and segregate cues in a way that is best explained by the BCI model, instead of the sensory segregation and forced fusion models (Cao et al., 2019). Furthermore, BCI efficiently manages the complexity of multisensory information and resolves the sensory ambiguity arising from probabilistic integration of prior knowledge with object features (Kersten, 2004). The brain frequently encounters incomplete or partially occluded object information requiring disambiguation through integration of multisensory cues and prior experience. This process generates reliable environmental predictions that support object categorisation, as described by the Predictive Coding (PC) framework (Friston, 2005; 2010).

Similar to the Bayesian frameworks, where predictions are updated on prior knowledge, the Predictive Coding (PC) framework generates and adapts predictions about sensory input aiming to minimising prediction errors. Particularly, the PC account posits that the brain maintains an internal model (i.e., priors or templates) of the environment which is constantly updated (Friston, 2010). This account implements the modulation of both bottom-up and top-down processes through an iterative process, with the former arising from the incongruity between the sensory information (i.e., forward connections) and the internal prediction, while the latter from the internal estimations built by the dynamic representations (i.e., backward connections). The top-down processes are guided by selective attention, which evaluate the most task-relevant information to be prioritised and processed. For instance, prioritising task-relevant stimuli might lead to a stronger representation of multisensory objects (Talsma, 2015).

Conclusively, in recent years, MI has been understood as consisting of a dynamic interplay between top-down and bottom-up processes. Moreover, researchers have found support for an optimal Bayesian integration of multisensory inputs, according to which the brain combines inputs from multiple senses according to their relative reliability.

1.2.2 Behavioural evidence of Multisensory Integration

To further assess the occurrence of MI, novel paradigms have been proposed to behaviourally observe it. More specifically, sensory illusions provide controlled conditions for examining how information from different modalities is integrated. For instance, when visual and auditory cues are conflicting, the brain may prioritise the input from one modality over the other, leading to a perceptual illusions such as the ventriloquist effect, where a sound is perceived as coming from a visual source, or reverse ventriloquism, where vision shifts the perceived location of a sound (Welch & Warren, 1980; Alais & Burr, 2004). Other well-known evidence for MI arises from the sound-induced flash illusion (SIFI; Shams et al., 2005; Hirst et al., 2020), demonstrating that auditory information can influence visual perception. This further supports the idea of integration

across multiple senses to interpret our environment, even at the cost of perceptual errors or illusions.

While incongruent stimuli like those in the SIFI might cause perceptual errors, congruent audiovisual stimuli have been shown to significantly enhance motor response speed and accuracy, particularly in older adults (Zou et al., 2023). Considering more high-level stimuli, multisensory enhancement (MSE) is reported across multiple experimental paradigms, in object recognition (Ernst et al., 2007), in cross-modal (audiovisual) categorisation tasks (Forster et al., 2002), in complex target detection (Molholm et al., 2004), and visual perception learning (Seitz et al., 2006). Further behavioural evidence arises from detection and localisation tasks in which performance is often facilitated (i.e., increased accuracy and direction location) by multisensory stimuli (Zampini et al., 2007; Tajadura-Jimenez et al., 2009; Sperdin et al., 2009). For instance, audio-somatosensory stimuli promote behavioural facilitation (i.e., faster responses) in multisensory trials during a detection task (Sperdin et al., 2009). Further support arises from cross-modal conflict studies: humans combine visual and haptic cues in a statistically optimal way (i.e., MLE) and the relative weight of each modality dynamically adjusts according to its reliability, leading to more precise perceptual estimates compared to those derived from either modality alone (Ernst & Banks, 2002). These findings suggest that the brain actively integrates sensory inputs based on probabilistic principles, further supporting the existence of a functional and adaptive multisensory system. Multisensory benefit has also been reported during a visuo-haptic grasping task as well: specifically, MI of size and position information leads to more effective gripping (i.e., faster actions with smaller grip apertures), with haptic position cues providing an increased benefit (Camponogara & Volcic, 2022). These findings demonstrate that in object perception, visual information might be fully integrated with haptic cues, including size and position, and that multisensory information supports effective gripping, as well as object perception.

1.2.3 Neurophysiological and Neuroimaging Evidence of Multisensory Integration

However, evidence for MI does not only come from behavioural studies, but also from neurophysiological and neuroimaging research. Specifically in the neurophysiological research, pioneering work by Stein & Meredith (1993) showed interactions between sensory inputs on neuronal activation from single-neuron level in the Superior Colliculus (SC) in cats (Stein & Meredith, 1993). More specifically, responses of the SC unit to bimodal stimuli exceeded the one to unimodal components. This finding built the foundation for the definition of multisensory enhancement (MSE) in SC. Moreover, the authors demonstrated that SC receives projections from multiple sensory cortical areas including the lateral supra sylvian visual area (LS) and the anterior ectosylvian area (AES), which further comprises an auditory (FAES) a visual (AEV) and a somatosensory area (SIV; Meredith & Stein, 1986; Wallace et al., 1992; Yau et al., 2015). Three main principles are proposed to regulate MI. Firstly, spatial and temporal congruency among a set of unisensory inputs create a stronger MI response. Secondly, the principle of inverse effectiveness postulates that stronger MI arises when the unisensory stimuli presented in isolation evoke weak responses (Stein & Meredith, 1993; Holmes & Spence, 2005; Starke et al., 2020; Stein et al., 2020). Lastly, both bottom-up and top-down processes influence MI in the mammalian brain (Choi et al., 2018). Bottom-up influences, such as stimuli strength, spatial location, and timing of sensory inputs, are considered: cross-modal projections among primary sensory cortices might mediate faster integration and supporting temporal integration of multisensory stimuli. For example, activation in the primary auditory cortex (A1) in response to visual input supports perceptual binding of temporally modulated sounds (Atilgan et al., 2018). Moreover top-down processes, such as learning, attention and locomotion, can modulate sensory responses across modalities, which in turn can alter the degree of MI (Choi et al., 2018).

With the emergence of neuroimaging technologies, the brain areas involved in MI have been identified. MI occurs both in primary sensory regions and association cortices (Ghazanfar & Schroeder, 2006). More specifically, early studies adopting low-level stimuli showed that silent lip

movements led to activation of the primary auditory cortex (A1; Calvert et al., 1997); basic auditory (i.e., tones and noises) and somatosensory (i.e., pulses to the fingertip) information activated the primary visual cortex (V1; Burton et al., 2002; Foxe et al., 2002; Röder et al., 2002). Beyond primary sensory areas, brain regions thought to process mainly visual object information, such as the Lateral Occipital Complex (LOC), showed cross-modal interactions when presented with simple, non-semantic auditory and visual stimuli (Molholm et al., 2004; Beauchamp, 2005). Overall, these findings show how the brain is fundamentally multisensory at various levels of its processing hierarchy. A key hub of MI is the posterior superior temporal sulcus (pSTS), which responds to somatosensory, auditory, and visual inputs, and shows enhanced responses to simultaneous auditory-tactile stimulation, demonstrating MI (Beauchamp et al., 2008). Considering complex object stimuli, a functional magnetic resonance imaging (fMRI) study found selectivity for man-made objects categories is correlated across vision and touch in the inferotemporal cortex (Pietrini et al., 2004), and that the somatosensory cortex is actively involved in processing purely visual information. While simply looking at but not touching an object, distinct patterns of activity emerge in S1 and the right S2 corresponding to the object's category (Smith & Goodale, 2015). These findings suggest that the brain automatically simulates the tactile properties and associated actions of visually perceived objects. Moreover, when objects are perceived haptically, category information is processed in S1 and LOC (Hernández-Pérez et al., 2017), further validating the role of LOC as a multisensory shape processing hub. Finally, multisensory object perception appears to be rapidly shaped by interactions between high-level semantic and early sensory areas: the anterior temporal and inferior prefrontal cortex showed audio-visual integration and sensitivity to semantic congruency within 50–100 ms, preceding responses in posterior occipital areas (Naci et al., 2012). Overall, neurophysiological and neuroimaging evidence support that the brain integrates sensory inputs via specialised multisensory regions both in humans and animals, with both top-down and bottom-up processes contributing to this integration.

Together, these findings support that perceptual and behavioural performance arise from integrated multisensory processing rather than isolated unisensory channels. However, whether multisensory integration is an automatic, stimulus-driven process in low-level sensory cortices (the ‘early’ account), or if it is a more cognitive, task-dependent processes that occur in higher-order association areas (the ‘late’ account) remains unanswered (Talsma et al., 2010).

1.2.4 The Early versus Late Multisensory Integration Debate

The early versus late debate on multisensory integration regards whether the brain combines information from different senses at the initial stages of sensory processing (early integration) or only after each sense has been processed individually in higher-order cognitive areas (late integration).

On one hand, evidence for early processing comes from both human and animal studies that identified the onset of MI (<100 ms) in cross-modal stimuli in primary sensory cortices (Foxe & Schroeder, 2005; Van der Burg et al., 2011), providing benefit in detection and localisation tasks. These early integrations lead to the cues’ co-occurrence in both spatial and temporal domains, also referred to as the ‘unity assumption’ (Ghazanfar & Schroeder, 2006; Talsma, 2015; Atilgan et al., 2018; Mercier & Cappe, 2020). Moreover, studies adopting electroencephalogram (EEG) showed that auditory-visual interactions occur early in humans (i.e., 90-95ms after stimulus onset; Cappe et al., 2010), suggesting that MI interactions involve the engagement of different configurations of intracranial generators, instead of one sense modulating the activity of another. On the other hand, late multisensory processing (> 250 ms) has been found in convergence zones to support binding (Beauchamp et al., 2004), attentional modulation (Donohue & Woldorff, 2011), semantic recognition and categorisation (Viganò et al., 2021). Moreover, evidence for integration in later stages arises from studies on reliability, congruency of multisensory stimuli and task relevance (Rohe & Noppeney, 2015; Beierholm et al., 2020; Cao et al., 2019; Rohe et al., 2019).

1.3 Categorisation of Unisensory Objects

The evidence presented thus far supports the view of multisensory organisation in the brain. To further understand MI, researchers have considered specific processes in which MI might be useful, for instance, object recognition. More specifically, when we encounter a dog we perceive it by its visual properties (e.g., shape, size, colour), the sounds it makes (e.g., barking), and its texture (e.g., by petting its fur). However, to understand how multisensory object perception works, one first needs to understand how each sensory system works in isolation to perceive and categorise objects.

1.3.1 Visual Object Categorisation

Categorisation is the fundamental cognitive ability by which we organise objects based on shared characteristics or properties, supporting efficient and fast identification and decision-making. However, the exact nature of the shared characteristics or properties on which categorisation relies generated a debate in the literature in the past century (Bruner et al., 1956; Rosch, 1973; Rosch & Mervis, 1975; Medin & Schaffer, 1978; Goldstone, 1994; Hahn & Ramscar, 2001).

One proposal suggests that categorisation leads to the identification of common features or similarities between objects, and that their categorisation is based on their function, features, similarities, or other universal criteria within the group (Su et al., 2020). For instance, an object is classified as a car if it has a defining set of features, such as four wheels, an enclosed cabin, and an engine for road transport. In fact, various influential models of categorisation are based on similarity, such as the Context Model (Medin & Schaffer, 1978) and Generalized Context Model (Nosofsky, 1986; Nosofsky, 1992). According to the Context Model, a small, sleek car might be classified as a 'sports car' because it shares key feature patterns in the sport car context, even if it also has features typical of family cars; while accordingly to the Generalized Context Model, an electric microcar or autonomous vehicle can be categorised as a 'car' due to its overall similarity

to stored exemplars in size, shape, and function, even though it differs in propulsion or control. Edelman (1999) found similarity-based models to be able to explain the visual categorisation of 3D objects, by representing them within a multidimensional similarity space in which category membership arising from proximity, which supports the view that similarity might be a useful heuristic for understanding how we recognise and categorise objects (Edelman, 1999). However, similarity-based approaches are challenged due to their vagueness and context-dependence (Murphy & Medin, 1985), as judgments regarding what is a 'typical car' can differ across countries, for example, a compact hatchback (e.g., a Fiat Panda) may be considered as standard in Italy or Europe but unusual in North America, and a crossover SUV may be considered as either a car or a truck in different regions.

To address this, other theories propose that categories might be structured around typical examples (i.e., prototypes) that are defined by their shared features and high within-category similarity (Rosch & Mervis, 1975). More specifically, cognitive studies have provided evidence of organisation of semantic information based on a detailed conceptual taxonomy (Rosch, 1999), formed to allow cognitive efficiency and organised in three levels of specificity: superordinate (general, e.g., 'vehicle'), basic (optimal recognition efficiency, e.g., 'car'), and subordinate (specific, e.g., 'Fiat Panda'). Furthermore, to understand the features that make a category and how the categorisation process works, several frameworks have been proposed. According to the Feature-based approach (Katz & Fodor, 1963), categories are defined through defining features, which are both necessary and sufficient, and all category members are equally representative (Katz & Fodor, 1963). However, this rigid boundary system fails to capture natural categories' fuzzy boundaries and variability. Thus, the influential Prototype Theory (Rosch, 1973) proposed less rigid boundaries in categorisation. Thus, according to this theory, categories are organised around central tendencies such as averaged representations of category members, referred to as prototypes (Rosch, 1973). More specifically, category membership is thought to rely on 'family resemblance' (Rosch, 1975), by which category members share characteristics or typical features,

thus accounting for typicality effects and flexible boundaries. Evidence for the Prototype Theory includes faster naming for typical members and efficiency at the basic categorisation level (Rosch & Mervis, 1975, Rosch et al., 1976). Support for this account arises from categorisation of patterns and geometric shapes (Posner & Keele, 1968; Homa et al, 1973; Reed, 1972), colour names (Berlin, 1969) and words (Labov, 1973; Lakoff, 1987), and from prototypes formation in memory (Hintzman, 1986; Posner & Keele, 1970; Neumann, 1974; Solso & McCarthy; 1981). However, Nosofsky and Palmieri criticised the Prototype Theory for oversimplifying complex categories, such as failing to account for atypical members, as well as a lack of specificity (Nosofsky & Palmeri, 1997), due to the high sensitivity of prototypes to context and individual prior experience (Gabora et al., 2008). Thus, the Exemplar Theory (Nosofsky, 1987; Nosofsky et al., 1989; Nosofsky & Palmeri, 1997) suggests categories are organised in collections of specific, stored instances (i.e., exemplars) rather than abstract prototypes, and that categorisation occurs through similarity comparison to retrieved exemplars. This approach better explains atypical member classification and category variability, with empirical evidence showing categorisation responses affected by exemplar similarity (Nosofsky, 1987), and better prediction of both accuracy and reaction times (Nosofsky, 1987; Nosofsky et al., 1989; Nosofsky & Palmeri, 1997) in comparison to the Prototype model. Finally, the Varying Abstraction Model (Vanpaemel & Storms, 2008; 2010) attempts to integrate both approaches, proposing a continuum from exemplar-based (minimal abstraction) to prototype-based (maximal abstraction) representations, with intermediate secondary prototypes formed through selective category member merging.

Though slightly different in their descriptions of features, the theoretical accounts described above agree that category formation relies on perceptual knowledge linking object categories to specific features or sensory properties. Expanding on this work, contemporary theories propose that categorisation involves two sequential steps: first, identifying relevant features and then learning ruling boundaries through exemplars or prototypes (Ashby & Maddox, 2004). However, whether this process is purely bottom-up (stimulus-driven) or incorporates top-down knowledge

(category-relevant features) is still unclear (Murphy & Allopenna, 1994; Heit, 1994; Ashby & Maddox, 2005; Riesenhuber, 2009; Seger & Miller, 2010; Rauss & Pourtois, 2013; Schendan & Ganis, 2015; Petkov & Petrova, 2019). Current evidence supports interactive models, where bottom-up sensory inputs activate primary cortical areas while top-down feedback from higher cortices refines perception through predictions and hypothesis testing (Rauss & Pourtois, 2013). For instance, the Competition between Verbal and Implicit Systems (COVIS) model (Ashby et al., 1998) proposes two different category learning systems. First, a rule-based (RB) system, which functions verbally and uses explicit hypothesis testing and logical rules while relying on semantic and working memory. Second, the information-integration (II) system, which functions implicitly, integrating multi-dimensional information without awareness through procedural learning and gradual improvement. At the neuronal level, these systems show dissociation: a frontal-based declarative system for explicit rule learning versus a basal ganglia-mediated procedural system (Cantwell et al., 2015; Ashby & Valentin, 2017). Evidence supporting the COVIS models stems from prototype distortion paradigm research, with behavioural studies revealing differential sensitivity to distortion and feedback (Casale & Ashby, 2008), and neuroimaging studies showing occipital cortex activation during visual category learning, with increased activation reflecting enhanced processing of discriminative visual features (Downing et al., 2001; Zeithamova et al., 2008). Moreover, behavioural evidence (such as categorical perception, CP) shows that the organisation of information into categories affects how objects are perceived: stimuli from different categories are more easily discriminated than equally distinct stimuli within the same category (Goldstone & Hendrickson, 2010).

Additionally, while learning novel categories, individuals adopt a continuum of categorisation strategies depending on data availability: with sufficient exemplars, the visual system uses an efficient feature comparison strategy against familiar category boundaries; while generalising from single examples, a more sophisticated generative model uses shape understanding to predict feature variability ranges (Morgenstern et al., 2019). Moreover, the

presentation format (i.e., blocked and interleaved) during category presentation affects attentional patterns during learning and generalisation of novel visual objects: categories learned with blocked presentation lead to equal attention to similarities and differences among objects, while interleaved presentation biases attention toward features that differed from the immediately preceding item (Carvalho and Goldstone, 2017).

1.3.1.1 Neural Models of Visual Object Perception: Recognition and Categorisation

Visual object processing follows a hierarchical structure from the primary visual cortex (V1) through specialised regions. Findings from clinical populations provide for two functional pathways: the ventral pathway (temporal lobe projection) supporting object recognition, and the dorsal pathway (parietal lobe projection) mediating action-related processing (Mishkin & Ungerleider, 1982; Jeannerod, 1988; Goodale & Milner, 1992). Moreover, the ventral pathway contains functionally specialised regions, including the Lateral Occipital Complex (LOC) for objects, the Fusiform Face Area (FFA) for faces (Kanwisher et al., 1997), and the Parahippocampal Place Area (PPA) for spatial processing (Zangaladze et al., 1999). More specifically, the LOC, located in the lateral fusiform gyrus, demonstrates robust activation for both 2D and 3D object representations (Grill-Spector et al., 2000). Research shows LOC specialises in hierarchical shape processing rather than low-level features (Malach et al., 1995), with posterior regions responding to simple fragments and anterior regions responding to complete objects (Grill-Spector et al., 2001). LOC neurons exhibit cue invariance, maintaining robust responses despite variations in size and position (Grill-Spector et al., 2001; Grill-Spector, 2003). However, the traditional category selectivity account was challenged due to the ventral pathway activation patterns reflecting elementary visual features rather than abstract categorical properties, suggesting a feature-based categorical coding where category representation partially depends on visual stimulus features (Bracci et al., 2017).

Nevertheless, categorical specificity appears to be a prevailing structural principle of the brain's sensory regions (Haxby et al., 2001). The temporal cortex processes object categories while the prefrontal cortex (PFC) handles higher-order categorical judgments (Jiang et al., 2018) in humans and monkeys (Kriegeskorte et al., 2008). Neurophysiological studies revealed category-selective neurons in monkey lateral prefrontal cortex (IPFC) during delayed match-to-category tasks: IPFC neurons showed increased responses to categorical information and maintained activity during delays, indicating working memory involvement; while inferior temporal (IT) neurons primarily analysed visual shape, with category selectivity potentially reflecting top-down PFC feedback (Freedman et al., 2001; 2002; 2003). The IPFC modulates sensory processing through top-down feedback, adjusting categorisation strategies based on context and experience. This prefrontal-temporal integration supports adaptive categorisation through continuous feedback loops between higher-order cognitive control and sensory processing regions.

1.3.2 Haptic Object Categorisation

Historically, the perception of objects has been primarily examined through visual inputs, reflecting a longstanding perspective that considers vision as the most complex and reliable sense (Hutmacher, 2019). However, visual information is not the only cue used for object categorisation, we also rely on haptic information. The sense of touch allows us to instantly perceive objects and to combine multiple features into percepts that support categorisation. For example, while searching for a phone in a bag, one might grasp a pack of tissues of similar shape, but quickly distinguish between the two based on material qualities perceived through touch. Unlike vision, which reconstructs a 3D representation from a 2D retinal image, haptic perception involves direct, exploratory contact with objects in three dimensions (Klatzky & Lederman, 2011). The haptic system can extract and process a wide range of object properties, including material substance and geometric form (Lederman & Klatzky, 1987; Kappers & Bergmann Tiest, 2013). Haptic perception integrates cutaneous inputs from skin mechanoreceptors and kinaesthetic inputs from tendons and muscles to extract geometric and material properties, including texture,

temperature, and shape (Lederman & Klatzky, 2009; Jones & Lederman, 2006). Overall, this indicates that touch enables object exploration and categorisation in different ways from vision.

In exploring how humans acquire tactile information, a central distinction should be made between active and passive touch, which differ in movement, source of sensory information, and the perceiver's role (Lawson et al., 2015; Ziat, 2022). Specifically, active touch, involving self-generated exploratory movements, proves more effective than passive touch for shape determination (Gibson, 1962), though passive touch provides valuable temporal information about motion and vibration (Mahns et al., 2006; Verrillo et al., 1969). Neuroimaging studies suggest a distinction in the neural correlates of active and passive touch (Simões-Franklin et al., 2011; Sumser et al., 2024), suggesting differing underlying processes. Passive touch significantly supports object categorisation by providing information on features such as texture and material, with temporal precision of tactile stimuli supporting discrimination based on the timing and sequence of inputs (Fleming et al., 2013; Fleming, 2017; Ziat, 2022; Mackevicius et al., 2012; Le et al., 2024). Moreover, research consistently shows that active haptic exploration enhances performance in recognition and discrimination tasks (Gibson, 1962; Lederman & Klatzky, 1987; 1990), and supports learning of unfamiliar objects (Newell et al., 2001; Lacey et al., 2009). Klatzky and Lederman's foundational research demonstrated humans achieve 96% accuracy in 3D object recognition within 3-5 seconds through active touch alone (Klatzky et al., 1985; Lederman & Klatzky, 1987; 1995). Further, they identified specific exploratory procedures: lateral motion for texture discrimination, contour following for precise shape extraction, and enclosure for structural information sufficient for basic-level identification (Lederman & Klatzky, 1987; Lacey & Sathian, 2022).

Furthermore, haptic perception provides unique viewpoints unavailable to vision, such as exploring object backs, with combined tactile-visual cues improving recognition accuracy (Newell et al., 2001; Edelman & Bülthoff, 1992). The salience of object properties varies by context: neutral

instructions prioritise hardness, texture, then shape, while haptic-focused instructions reverse this sequence, and visual-haptic tasks emphasise shape dominance (Klatzky et al., 1987). This demonstrates an adaptability in haptic object perception. Furthermore, haptic processing includes ‘container haptics’ for inferring enclosed objects without direct contact (Frissen et al., 2023). Recognition shows similar size rescaling processes across visual and haptic modalities (Craddock & Lawson, 2009) but exhibits orientation sensitivity, with decreased performance for non-canonical views reflecting biomechanical and exploratory constraints of hand movement (Lawson, 2009; 2011; Lawson & Bracken, 2011).

Based on this haptic specificity, we can recognise and categorise objects through haptic information alone. At the basic level of objects categorisation (e.g., distinguishing a cup from a plate), shape information is important (Schwarzer, Kufer, & Wilkening, 1999), and the haptic system is an expert in identifying shapes (Klatzky et al., 1985). Despite the challenges in extracting precise geometric data, approximate shape information, combined with texture and hardness, is often sufficient for accurate categorisation (Klatzky & Lederman, 2011). Lederman and Klatzky identified a two-stage exploration sequence: initial ‘grasp-and-lift’ movements followed by specialised patterns targeting diagnostic object features (Lederman & Klatzky, 1990), as well for toy objects (Haag, 2011). Research demonstrates that haptic systems form realistic perceptual spaces correlating strongly with visual ones (Cooke et al., 2007; Gaissert et al., 2010), with both modalities relying primarily on shape features when categorising natural objects (Gaissert & Wallraven, 2012). Studies using parametrically controlled novel stimuli reveal that haptic and visual modalities create unified perceptual spaces incorporating shape and texture dimensions (Cooke et al., 2007): while vision prioritises shape over texture, haptic perception shows more balanced feature weighting. Gaißert, et al., (2011) further extended these findings by showing that both visual and haptic modalities form similar perceptual spaces for novel 3D objects (Gaißert et al., 2011), reinforcing the view that haptic processing is suitable and robust for categorisation based on intrinsic object properties. Additionally, categorical perception, traditionally studied in

vision, extends to haptic shape perception, with category learning reshaping haptic representations and improving cross-boundary discrimination (Gaißert et al., 2012). Overall, these findings demonstrate touch's ability for independent category structure learning and efficient boundary identification (Knight et al., 2014).

1.3.2.1 Neural models of haptic objects: recognition and categorisation

Haptic object recognition involves a distributed neural network processing tactile information from primary somatosensory cortex (S1) in the postcentral gyrus through higher-order cortical areas (Klatzky et al., 1985; Lederman & Klatzky, 1990). Somatosensory information segregates into ventral and dorsal pathways: the ventral pathway through parietal opercular cortex processes texture (Mishkin, 1979; Kitada et al. 2005; Simões-Franklin et al. 2011), while the dorsal pathway transmits spatial information to the intraparietal sulcus (IPS) and frontal eye fields, integrating tactile inputs with object orientation and hand movements (Sathian et al., 2011; Sathian, 2016). The IPS serves as a key hub for haptic shape processing, with anterior IPS (aIPS) involved in object size and shape processing (Perini et al., 2020). Notably, the LOC, traditionally associated with visual processing, shows robust activation during haptic exploration even without visual input (Amedi et al., 2001; 2002; James et al., 2002). Overall, these cross-modal activations support modality-independent object shape representations, suggesting that both tactile and visual information converge in higher-order cortical areas, allowing the brain to form robust, multisensory object representations. Moreover, the interaction between the IPS and LOC is thought to be important for building an abstract object representation, integrating its spatial and geometric properties (Lacey & Sathian, 2014). Categorical information is encoded across multiple regions: multivariate pattern analysis reveals category decoding from LOC, somatosensory, and parietal cortices (Hernández-Pérez et al., 2017). Hierarchical processing shows stimulus-selective patterns in precentral and postcentral gyri, with category-selective responses emerging in prefrontal cortex (Malone et al., 2019). The network extends to visual motion areas, with the

medial superior temporal area (MST) responding to vibrotactile stimulation (Beauchamp et al., 2007), indicating integration across traditionally separate sensory systems.

1.3.3 Auditory object Categorisation

As haptic exploration allows for efficient objects categorisation based on tactile cues alone, auditory processing similarly supports object categorisation through sound-based features. Auditory perception involves processing complex acoustic environments through Auditory Scene Analysis (ASA), which groups acoustic components into coherent 'auditory objects' using sequential and simultaneous integration principles (Bregman, 1993; 1984; 2017). The precise nature of an auditory object, whether it is a 'stream' of inputs, a 'mental event,' or a 'region in time-frequency space', remains a subject of academic debate (Kubovy & Van Valkenburg, 2001; Griffiths & Warren, 2004; O'Callaghan, 2008; Bullot & Égré, 2010; Bizley & Cohen, 2013; Bizley et al., 2016). These objects can be defined as perceptual constructs reflecting the auditory system's interpretation of acoustic events, characterised by spectro-temporal regularities and invariance to contextual variations (Griffiths & Warren, 2004; Bizley & Cohen, 2013). Moreover, auditory object recognition relies on multidimensional acoustic cues, including pitch, timbre, loudness, and temporal structure (McAdams, 1993). It involves two-stage processing: grouping sounds into 'proto-objects,' followed by competitive selection of coherent auditory streams (Bregman, 1990). Importantly, the auditory system demonstrates invariance to spectro-temporal changes, maintaining stable object recognition despite acoustic variability (Heald et al., 2017; Town et al., 2018).

Auditory categorical perception manifests as abrupt perceptual shifts despite continuous acoustic changes, exemplified by phoneme boundary effects (Liberman et al., 1967; Kuhl & Padden, 1982). These boundaries exhibit 'perceptual plasticity', adapting based on context, experience, and statistical regularities in the acoustic environment (Goldstone, 1998; Heald et al., 2017). The system extracts hierarchical information from systematic regularities to disambiguate

ambiguous signals (Feldman, 2000). Beyond speech, research has also investigated the perception of ‘object sounds’ generated from both animate (e.g., human non-vocalisations, animal sounds) and inanimate (e.g., environmental or man-made machinery) sources, showing that such sounds are categorised and represented as auditory objects within cortical networks, involving the superior temporal cortex, anterior temporal and prefrontal areas (Brefczynski-Lewis & Lewis, 2017). Moreover, research on everyday environmental sound categories reveals that they are primarily organised by semantic meaning rather than acoustic similarity, such that basic-level categories (e.g., car, dog bark) are most salient, with organisation based on sound source (natural vs. man-made) and semantic meaning (Guastavino, 2007). For instance, individuals can freely group sounds into high-level semantic categories such as ‘human’ ‘animal’ or ‘mechanical’ (Bones, et al., 2018), supporting the existence of taxonomy over raw acoustic properties. Furthermore, the perceptual magnet effect, originally demonstrated in speech (Kuhl, 1991), extends to non-speech domains where sounds acoustically close to category prototypes are harder to distinguish than those further away (Guenther & Gjaja, 1996; Lotto et al., 1999). Within-category versus between-category effects show faster, more accurate discrimination between different categories than within the same category (Liberman et al., 1957; Howard & Ballas, 1980). Harmonic content serves as a key differentiating attribute, particularly for distinguishing vocalisations from non-vocalisations (Brefczynski-Lewis & Lewis, 2017). Furthermore, auditory category learning demonstrates robust flexibility across training methods, with feedback-based training proving most efficient (Obasih et al., 2023), suggesting a robust and flexible learning mechanism in the brain. This process involves interplay between unsupervised statistical learning, forming initial sensory maps, and reinforcement learning, refining representations through feedback (McMurray & Chiu, 2020). The brain first learns the structure of the world through passive observation and then uses feedback loops to attach meaning and behavioural importance to that structure. The principle of meaning-driven, context-sensitive categorisation (Dubois, 2000) extends from individual sounds to complex auditory scenes and soundscapes (Dubois, 2000). Soundscape

perception integrates acoustic cues with semantic knowledge (Woodcock et al., 2017), with categorisation influenced by functional roles (e.g., ‘dialogue,’ ‘background music’) and emotional interpretation. Research shows that soundscapes reflecting human activity are perceived as more pleasant than mechanical-dominated environments (Dubois et al., 2006; Davies et al., 2013).

1.3.3.1 Neural models of auditory objects: recognition and categorisation

The auditory system employs hierarchical processing with dual-pathway organisation: a ventral ‘what’ stream extending from primary auditory cortex (A1) through anterior temporal lobe to ventrolateral prefrontal cortex for object identification, and a dorsal ‘where/how’ stream projecting to posterior parietal cortex for spatial localisation (Tsunada & Cohen, 2014; Bizley & Cohen, 2013; Shamma & Micheyl, 2010). Hierarchical feature extraction progresses from simple frequency and intensity coding in A1 to complex, behaviourally relevant sound processing in higher-order areas. Temporal coherence mechanisms bind acoustic features into unified objects through correlated neural activity across frequency channels (Micheyl et al., 2013; Liu et al., 2017; Rezaeizadeh & Shamma, 2021). Electrical neuroimaging reveals rapid object discrimination within 100 milliseconds, with living versus non-living sound categories activating distinct cortical networks (Lewis et al., 2006; Engel et al., 2009). Two competing models explain auditory categorisation: the two-stage hierarchical model proposing sequential auditory cortex feature learning followed by prefrontal categorisation decisions (Meyers et al., 2009; Jiang et al., 2018), and the Dual Learning System (DLS) model featuring parallel ‘sound-to-rule’ (explicit, cortico-thalamic-basal ganglia) and ‘sound-to-reward’ (implicit, auditory-motor-striatal) systems (Chandrasekaran et al., 2014). The COVIS framework extends this to auditory domains, with rule-based learning mediated by frontal-striatal circuits and information-integration learning involving basal ganglia procedural mechanisms (Ashby & Spiering, 2004; Ashby & Valentin, 2017). Moreover, category learning induces cortical reorganisation, with MEG studies revealing that information-integration learning activates auditory sensory-motor pathways (superior temporal gyrus, inferior frontal gyrus), while rule-based learning engages distributed cognitive control

networks (Feng et al., 2021). The striatum and superior temporal sulcus support implicit category learning from naturalistic environments (Lim et al., 2019). Beyond basic perception, the neural system can generalise auditory learning, extending patterns to the categorisation of novel exemplars (Heald et al., 2022).

Overall, the research reviewed indicates that object categorisation occurs across multiple sensory modalities, each with specific strengths and limitations. For example, considering visual and haptic modalities, while vision provides rapid, detailed information but cannot access hidden surfaces, whereas haptic exploration enables full 3D perception, although more slowly and effortfully. Moreover, while haptic exploration is specialised at perceiving texture and surface properties, auditory perception supports efficient localisation of stimulus origin. Together, these complementary competences highlight how the senses might work together to achieve coherent object representation and efficient categorisation.

1.4 Categorisation of Multisensory Object

In the real world, objects are rarely encountered through a single sensory modality. For instance, when seeing a dog while hearing its bark (Chen & Spence, 2010). In our environment, we can often use cues across different modalities, resulting in a more complex but also ecologically valid consideration of multisensory object categorisation. The following section, therefore, examines how the brain integrates information from different sensory modalities to create a unified and robust representation of objects, leading to enhanced perception and more efficient categorisation.

1.4.1 Visuo-auditory interactions in object perception and categorisation

Audio-visual objects represent perceptual constructs formed through the integration of multiple stimulus properties within the brain (Bizley et al., 2016). Research has consistently demonstrated that multisensory exposure affects subsequent unimodal processing. In a series of studies, Murray and colleagues established that single-trial audio-visual exposure significantly

improves auditory and visual object discrimination compared to unimodal exposure alone (Lehmann & Murray, 2005). The authors adopted ecologically valid paradigms resembling the real-world multisensory processing, such as continuous old-new recognition task adopting objects both animate and inanimate. The authors replicated this effect across behavioural, electrophysiological and neuroimaging studies (de Meo et al., 2015; Lehmann & Murray, 2005; Matusz et al., 2017, 2019; Thelen & Murray 2013, Murray et al., 2018; Murray & Sperdin, 2010; Tovar et al., 2020). This multisensory advantage manifests as enhanced discrimination accuracy across both visual and auditory domains, though response speed remains unaffected. Critically, semantic congruency modulates these effects: congruent audio-visual pairings facilitate subsequent unisensory recognition, while incongruent pairings impair performance (Thelen et al., 2015). These findings show how the initial context during the encoding stage affects both subsequent processing and discrimination of the identical sensory trace: perceiving a congruent audio-visual object aids recollection of the same object when presented visually and auditorily only. Further, neuroimaging evidence reported that congruent multisensory presentations produce stronger LOC activation during visual recognition tasks, whereas incongruent pairings activate differentially distinct networks, such as weak activation in the right LOC and strong activation in the posterior STS (Matusz et al., 2019).

This interplay between sensory modalities extends beyond memory improvement and has been shown to be highly dependent on the contextual environment and the relationship between the stimuli. Specifically, semantic relationships between co-occurring stimuli serve as fundamental modulators of multisensory integration. In a series of experiments, Chen & Spence (2010) showed that semantically congruent audio-visual stimuli improve visual identification, while incongruent pairings impair identification. This benefit is proposed to occur through rapid semantic integration within a short-term buffer that forms coherent multisensory object representations, which in turn facilitates subsequent perceptual tasks as identification. Moreover, given that auditory perception shows 360° coverage compared to vision's limited 135° field, the auditory context frequently

provides initial perceptual context for what we are about to see (visual processing). Recent studies extend these principles to dynamic scenarios during the categorisation of ambiguous visual objects. Williams et al. (2022) found that naturalistic sounds paired with ambiguous visual stimuli bias object identification toward auditorily-congruent interpretations, indicating that ongoing audio-visual integration effects on novel stimulus processing (Williams et al., 2022). However, these findings challenge views positing that semantic knowledge is fundamental for visual perception (Enge et al., 2023). While consistent with low-level, implicit probabilistic inference processes, which suggest that visual perception and semantic, top-down goals rely on distinct mechanisms, these results highlight the critical role of multisensory integration in shaping visual object representation and categorisation. Furthermore, multisensory audio-visual benefits extend to perception of moving objects, where congruent sounds facilitate visual object detection during self-motion in younger adults (Roudaia et al., 2018), with cross-modal feature-based attention enhancing visual motion pattern recognition (Hanada et al., 2019).

1.4.1.1 Visuo-auditory interactions in the perception of novel objects

Multisensory categorisation processes engage distinct learning systems, including rule-based and information-integration categorisation frameworks (Ashby & Maddox, 2005). Thus, knowledge acquired through multisensory training can be transferred to facilitate categorisation of novel, untrained stimuli, demonstrating the acquisition of abstract, cross-modal category knowledge. The work of Wu et al. (2021) provides strong evidence for this, showing that multisensory training improves categorisation accuracy for both trained and novel stimuli, particularly under ambiguous conditions. Participants trained on car images and sounds showed enhanced performance when congruent cross-modal information was available in another modality (e.g., a matching car sound for a car image), with benefits most pronounced for untrained, high-distortion stimuli (Wu et al., 2021). This supports the view that the brain acquires abstract, cross-modal category knowledge that can be flexibly applied to novel exemplars, confirming the principle of inverse effectiveness: the more degraded the unisensory stimulus, the

more robust the benefit from multisensory integration. Beyond the perceptual benefits for audio-visual stimuli, another complex question is how the brain learns to associate multisensory objects with abstract symbols or names. Investigating this, Viganò et al. (2021) mapped the temporal course of symbolic categorisation of animated novel audio-visual objects in a delayed-match-to-category-name task. The authors found that the neural mechanisms underlying symbolic categorisation of multisensory objects follow a hierarchical three-stage representational processing (Viganò et al., 2021). Initially, sensory cortices exhibit enhanced neural selectivity for relevant features following training, with the frontoparietal network processing multisensory object-specific responses. Evidence arises from high sensory segregation and sharpened neural selectivity in the auditory and visual cortices after training. The second stage involves the left angular gyrus, which integrates sensory inputs into unified object representations and links them to corresponding linguistic symbols. Local and distributed coding for audio-visual object identities emerges post-learning in the left angular gyrus, left middle, and right inferior frontal gyrus, demonstrating multisensory integration during symbolic categorisation. The final stage involves the hippocampus, which encodes abstract, format-invariant representations that generalise across both multisensory objects and category names. At this stage the categories are represented regardless of their initial presentation format (i.e., whether through words or as objects). Evidence for this last stage comes from shared representational codes for categories and objects in associative cortices and the left angular gyrus, which indicate a cross-modal integration format invariant process. Finally, the right hippocampus revealed generalisation across both object and word presentations, supporting abstract categorical representation formation. These findings support a clear hierarchical progression, from sensory-specific processing through integrated multisensory representation to abstract symbolic thought. In particular, the format-invariant coding found in the hippocampus in the post-learning provides evidence for how the brain forms high-level abstract concepts from multisensory experiences, with distinct brain

regions contributing to different levels of abstraction (i.e. differential generalisation accuracy in the right hippocampus and not for the left angular gyrus, Viganò et al., 2021).

1.4.2 Visuo-haptic interactions in object perception and categorisation

In addition to audition, the sense of touch provides a valuable source of information, which is often integrated with vision to guide object perception and interaction. Cross-modal visuo-haptic object recognition exhibits systematic asymmetries, with visual-to-haptic transfer typically demonstrating more accurate speed and accuracy compared to haptic-to-visual recognition (Easton et al., 1997; Lacey & Sathian, 2014). This asymmetry manifests particularly in viewpoint-dependent recognition tasks, where recognition accuracy was higher when visual perception preceded haptic encoding, particularly with orientation changes (Newell et al., 2001) and was suggested might be due to complementary exploratory patterns (i.e., hands exploring from the back and eyes from the front). The underlying mechanism of this asymmetry involves differential cognitive costs: haptic recognition is profoundly impaired by orientation changes, whereas visual recognition efficiently handles viewpoint normalisation through mental rotation capabilities (Lawson, 2009; 2014). One possible explanation for this asymmetry might arise from modality-specific encoding strategies: while vision prioritises global shape information, haptics weights local features and texture equally with shape (Cooke et al., 2007). However, top-down cognitive control significantly modulates these encoding processes. Accessing and comparing information across modalities likely depends on the strategic allocation of attention. Lacey and Sathian (2020) showed that when participants anticipate cross-modal testing, visual encoding strategies adapt through longer, more diffuse fixation patterns, suggesting deliberate construction of robust object representations (Lacey & Sathian, 2020). Similarly, haptic exploratory patterns flexibly adjust based on task demands and object properties, demonstrating goal-directed information gathering across both modalities (Leo et al., 2022; Sciutti et al., 2019). The existence of such flexible, goal-directed encoding challenges the idea that the modalities are completely separate: if both senses can adapt for a common goal, perhaps they contribute to a shared representation. Therefore, the

shared multisensory representation that has been suggested for cross-modal recognition might extend cross-modal categorisation. This is supported by evidence of symmetric cross-modal transfer of category structure following unimodal (visual or haptic) category learning, with performance improvements generalising to novel objects across modalities (Wallraven et al., 2014). This seamless transfer across modalities suggests unified visuo-haptic shape representations rather than separate modality-dependent knowledge systems. The neural basis for this integration is suggested to involve haptic exploration triggering top-down visual imagery through prefrontal cortex projections to the LOC, facilitated by spatial processing in the IPS (Lacey & Sathian, 2014). The IPS enables whole-from-part object reconstruction for both familiar and unfamiliar objects, allowing visual representation reconstruction from haptic input and providing neurobiological support for unified multisensory object representations.

1.4.2.1 Visuo-haptic interactions in object perception of novel objects

The perception of novel objects requires integration across visual and haptic systems despite their distinct specialisations. Multidimensional scaling analyses reveal modality-specific perceptual spaces during visuo-haptic perception and categorisation of 3D novel objects: vision prioritises macro-geometric features (shape), while haptics weights both shape and micro-geometric properties (texture) equally (Cooke et al., 2006; 2007). Despite these differential weightings among the visual and haptic systems, behavioural evidence shows successful cross-modal integration, with efficient category information transfer between modalities (Yildirim & Jacobson, 2013). This integration incorporates orientation-specific information through consistent exploration patterns across cross-modal tasks, as participants use consistent exploration patterns in cross-modal tasks even if it results in less efficient recognition (i.e., less accurate; Ernst et al., 2007). These findings suggest that the brain is actively working to unify information from different sensory modality, instead of processing them in isolation. Neural mechanisms underlying this visuo-haptic integration involve neural networks, such as the LOC and IPS, being involved in visual-haptic processing of object shape and the left anterior fusiform gyrus (laFG) showing activation

during both visual and haptic shape discrimination, with multisensory enhancement effects being most pronounced for salient stimuli (Kim & James, 2010). These findings support neuronal convergence models of haptic and visual sensory information processing. This is strongly corroborated by computational models. The MVH model demonstrates that multisensory representations of prototypical 3D shapes emerge through sensory-specific forward projections critical for category learning and transfer (Yildirim & Jacobson, 2013). Similarly, deep neural network structures reveal that bimodal visuo-haptic training produces more abstract, robust, and viewpoint-invariant object representations compared to unimodal training (Jacobs & Xu, 2019). These convergent findings support multisensory facilitation during perceptual category learning and provide evidence for shared visuo-haptic object representations that overcame modality-specific processing limitations.

1.4.3 Audio-haptic interactions in object perception and categorisation

When exploring a surface, we often perceive its texture as well. Texture perception operates through multisensory mechanisms involving both tactile and auditory processing systems. The Duplex Theory proposes that texture perception relies on dual coding: fine textures are processed via high-frequency vibrations detected by Pacinian corpuscles, while coarse textures utilise spatial cues from slowly adapting mechanoreceptors (Klatz, 1989; Hollins and Risner, 2000). The link to audition lies in the vibrational input: the same frequency information that underlies tactile roughness is also a fundamental component of auditory pitch and timbre. This shared reliance on frequency information suggests a connection between the two senses. Neuroimaging evidence revealed that vibrotactile exploration activates both somatosensory (S1, S2) and auditory cortical regions, with A1 activation modulated by stimulus movement (Roberts et al., 2023). These findings suggest that vibration arising from sliding haptic exploration led to multisensory cortical activation in the processing of roughness. This neural connectivity produces behavioural effects: congruent audio-haptic combinations improve texture judgment accuracy and slow down exploratory hand movements, suggesting integrated perceptual processing (Montano-Murillo et al., 2023).

Importantly, transcranial magnetic stimulation of the auditory cortex impairs tactile roughness discrimination in the absence of sound, establishing causal involvement of auditory regions in purely haptic texture processing (Yau et al., 2009). Together, these findings suggest that the brain recruits neural substrates typically associated with hearing to make sense of what we feel, and thus, audio-haptic sensory combinations are important for ecologically valid and embodied object perception experience.

From a perceptual viewpoint, psychophysical evidence reveals shared computational principles across audio-haptic modalities: subjective roughness perception follows identical psychophysical curves for auditory and haptic stimuli, with roughness judgments peaking at equivalent frequency ratios regardless of presentation modality (Bernard et al., 2022). This perceptual correspondence supports common underlying mechanisms for processing temporal fluctuations across these sensory systems. Specifically, the left auditory cortex is activated for frequencies that can be perceived by the human ear, but not for those below the perceptual threshold (Nordmark et al., 2012). This supports the meta-modal brain organisation theory (Pascual-Leone & Hamilton, 2001), which posits that brain areas support sensory processing by executing specific computations, irrespective of the initial input modality. This view could be further extended to an inherent multisensory nature of object perception and representation (Pascual-Leone & Hamilton, 2001; Tal et al., 2016). Supporting this view, evidence from neuroimaging studies shows how different sound frequencies activate distinct somatosensory cortex regions, with activation patterns correlating with perceptual judgments (Pérez-Bellido et al., 2018). Overall, these results are relevant for object categorisation since assessing an object's surface roughness may involve both vibrations and sounds arising from the interaction with the object (Yau et al., 2009). For object texture processing, frequencies in the 100-300 Hz range are crucial (Manfredi et al. 2014), with audio-tactile associations developing through corresponding highly frequency-specific sound and vibration inputs, in order to facilitate effective object perception and categorisation. Taken together, these findings show that the texture perception is

an inherently multisensory process, in which the brain utilises the auditory cortex's expertise in frequency analysis to support and refine judgments of tactile roughness, creating a unified audio-haptic percept built on shared neural mechanisms.

Overall, the evidence discusses demonstrated robust interactions among audio-visual, visuo-haptic, and audio-haptic modalities. Across these pairings, we observed more efficient object categorisation, supported by converging behavioural data and neuroimaging evidence. In particular, overlapping activation in multisensory brain regions emerged for visuo-haptic and audio-haptic conditions, underscoring shared neural mechanisms.

1.5 Outline and Scope of this Thesis

Taken together, the literature discussed thus far demonstrate the category specificity organisation in the brain across the visual (Rosch, 1978; Peelen & Downing, 2017), auditory (Tsunada & Cohen, 2014; Doehrmann et al., 2008; Feng et al., 2021; Brunel et al., 2013), and tactile (James et al., 2007; Lacey et al., 2009; Yildirim & Jacobs, 2013) domains. In addition, the literature further supports multiple features integration for both differentiation between categories and generalisation to novel exemplars in object category learning (Goldstone & Hendrickson, 2010; Pérez-Gay et al., 2017). Although, an abundance of evidence for multisensory interactions in the brain exists (Ghazanfar & Schroeder, 2006; Alais et al., 2010), to date, the processes underlying multisensory cues integration in categorisation and generalisation remain poorly understood (Newell et al., 2023). To address this gap in knowledge, the present thesis aims to investigate the role of multisensory perception in the formation of object categories. Specifically, this thesis explores the contribution of visual, auditory and tactile information in the formation of multisensory perceptual categories. To do so, Chapter 2 examined whether prior multisensory (audio-visual) exposure influences the subsequent categorical perception of familiar objects. Chapter 3 progresses to investigate whether motion and sound cues affect the perceived shape of unfamiliar objects using judgements of inter-objects similarity. Moreover, Chapter 4

investigated whether tactile and visual object motion cues influenced the categorisation of unfamiliar object shapes, and if the formation of these multisensory categories subsequently benefited generalisation to novel exemplars. Finally, Chapter 5 investigates how the perceptual system resolves visuo-haptic conflict in a categorisation task in virtual reality. The theoretical implications of the findings from the behavioural experiments reported in this thesis are discussed in detail in Chapter 6, along with direction for future research.

Chapter 2

Audio-Visual Categorical Perception of Familiar Objects

2 Audio-Visual Categorical Perception of Familiar Objects

2.1 Abstract

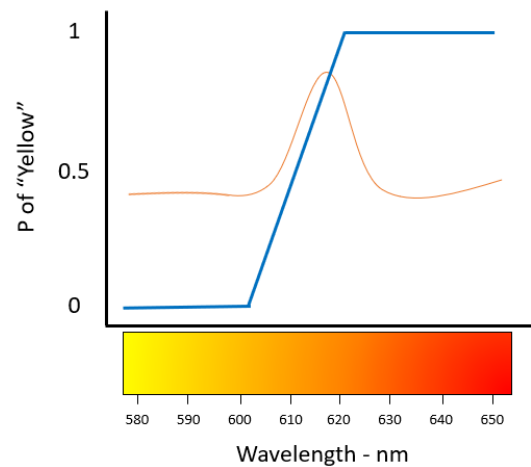
Every day we recognise and categorise countless objects, often using information from multiple sensory modalities, such as vision and audition. For example, when seeing a dog, we often hear its bark (Chen & Spence, 2010). However, most of the investigations to date on how we categorise objects has focused on single one unimodal sensory modality, mainly vision. Thus, the processes underlying multisensory object categorisation are still unclear. The aim of the experiments presented in this Chapter is to investigate whether prior audio-visual multisensory exposure modulates the subsequent visual categorical perception of familiar objects. To do so, we developed a novel two-phase paradigm involving bimodal learning followed by unimodal visual discrimination and categorisation tasks on morphed object continua. In Experiment 1A, we confirmed that auditory stimuli were semantically congruent with their corresponding visual object categories. In Experiment 1C, we assessed the perceived similarity for familiar visual objects, created along a morphed continuum. Although, Experiment 1B revealed no multisensory benefit on the category boundary, participants consistently categorised objects, particularly within-category items, as belonging to distinct categories, even without evidence of increased discriminability at the category boundary. These findings suggest that while auditory cues can be semantically aligned with visual objects, prior multisensory experience may not enhance categorical perception or modulate category boundaries, suggesting a more complex interplay between sensory modalities in object categorisation.

2.2 Introduction

Object recognition is important for our survival and is defined as the ability to both identify objects with precision and to categorise them as part of a specific category (DiCarlo et al., 2012). Research of human perceptual processing have revealed that the categorical organisation of information has a direct effect on how objects are perceived (Beale & Keil, 1995; Harnad, 1987; 2005; Goldstone & Hendrickson, 2010). A phenomenon that has a role in how individuals discriminate and categorise stimuli in their environment is categorical perception (CP), which refers to the tendency for continuous sensory information for to be perceived as belonging to distinct categories instead of as points along a seamless continuum. Specifically, this phenomenon is the characterised by enhanced discrimination between stimuli that straddle a category boundary and worsen discrimination for stimuli within the same category, even when the physical distance between the stimuli remains constant (Harnad, 1987; Goldstone & Hendrickson, 2010). The category boundary is defined by the categorisation response function, predicting the increased (or peak) performance during the discrimination task. This behavioural phenomenon emphasises the importance of categorical boundaries in shaping our perceptual experience, and how perceptual systems organise incoming information into discrete categories, thus facilitating more efficient recognition and decision making. Figure 1 shows the different patterns observed in both categorisation and discrimination tasks on a continuous stimuli space, i.e., wavelength going from bright red to yellow. For instance, two different shades of red around 650 nm are less discriminable compared to red and orange around 610 nm, despite the physical difference between the two being the same.

Figure 1

An illustration of the Categorical Perception phenomenon.



Note: The plots report the proportion of correct responses on the y-axis, across the continuous stimulus space on the x-axes. Categorisation responses are reported in blue, and discrimination responses in orange. Categorical perception is typically demonstrated by an increase in discriminability to stimuli that lie across the participant's perceived category boundary.

The delimitation of categories acts as a cognitive scaffold, shaping how we differentiate and interpret the numerous stimuli encountered in daily life. Categorical perception tends to occur for highly similar stimuli that shows subtle differences, suggesting that the visual system is attuned to discriminate among objects sharing many features, but that are also slightly different. This phenomenon develops as a consequence of learning to discriminate between highly similar objects (Gauthier et al., 1999; Wong et al., 2009), whereby repeated exposure and experience refine perceptual boundaries and strengthen the categorisation of stimuli. For instance, discriminating among exemplars within a category (e.g., different types of glasses) is perceptually more demanding than categorising across distinct object classes (e.g., a glass versus a bottle) due to their overlapping feature space and shared structural components (Op de Beeck et al., 2003). Thus, when a stimulus crosses a category boundary, a finer discrimination is required, as these stimuli have features that bridge different conceptual groupings. In contrast, stimuli within the same category tend to share overlapping features, thereby reducing the demand for fine-grained

perceptual differentiation. This phenomenon challenges continuous models of perception rooted in early psychophysical theories (Fechner, 1948), assuming that perceptual experience maps linearly onto physical stimulus properties. Instead, categorical perception suggests that higher-level cognitive processes, such as conceptual knowledge, can modulate perceptual discrimination, especially near category boundaries (Goldstone, 1994b; Livingston et al., 1998), providing insight into the nature of object representations. Evidence of categorical perception has been found for basic perceptual dimensions such as colour and phonemes (Briscoe, 2020; Kronrod et al., 2016), as well as for complex dimensions such as familiar objects (Newell & Bülthoff, 2002), faces (Beale & Keil, 1995), and the sex of faces (Bülthoff & Newell, 2004).

Although categorical perception focuses on how sharply and consistently boundaries are defined between categories, another important aspect of category learning is the underlying perceptual similarity structure of the stimuli themselves. Perceived similarity has long been proposed as a central mechanism underlying category learning and generalisation (Shepard, 1987; Goldstone, 1994b; Ashby & Maddox, 2005). Thus, categorisation performance is influenced by the psychological distance between exemplars: the more similar a stimulus is to a known category member, the more likely it is to be categorised accordingly. Particularly, similarity-based models of category learning suggest that new stimuli are classified by comparing them to stored exemplars or prototypes based on their psychological distance (Shepard, 1987; Nosofsky, 1986), and thus categorisation decisions are guided by how perceptually similar a stimulus is to known category member.

Recent findings have shown that object recognition benefits from the integration of information from multiple sensory systems: cross-modal stimulation (e.g., simultaneous visual and auditory input) can benefit perception as measured by reaction time performance (e.g., faster reaction time; Forster et al., 2002), complex target detection (Molholm et al., 2004), and perceptual learning (Seitz et al., 2006). Therefore, stimuli presented in different sensory

modalities offers complementary information about objects, thereby benefitting identification performance. In a series of multidisciplinary studies, Murray and colleagues examined the impact of single-trial multisensory exposure on subsequent object discrimination, finding evidence for a multisensory benefit for object perception (Lehmann & Murray, 2005; Matusz et al., 2019; Thelen & Murray, 2013; Murray & Sperdin, 2010; Tovar et al., 2020). Specifically, in a behavioural study adopting a continuous recognition task for animate/inanimate objects, Lehmann & Murray (2005) report more accurate object discrimination for stimuli previously experienced through audio-visual semantically congruent exposure (i.e., a cat meowing), compared to unimodal or semantically incongruent exposure (Lehmann & Murray, 2005). This task, which was designed to mimic real-world scenarios and investigate different types of memory information (i.e., episodic and semantic), provided a tool for understanding how the integration of audio-visual information affects subsequent object discrimination. This finding suggests that multisensory memory traces are available at the stage of perceptual processing. Electrophysiological and neuroimaging studies (i.e., event-related potentials – ERPs) investigated both the temporal dynamics and spatial localisation of this multisensory benefit, reporting that the left LOC showed greater BOLD activation and was most responsive at 60–136 ms following stimulus onset, only if the target visual stimulus was previously experienced in a multisensory context (Murray et al., 2004; 2005). Overall, these findings suggest that single exposure to multisensory object stimuli can influence subsequent perceptual discrimination and categorisation processes. However, some research has failed to detect changes in perceptual performance or neural signatures following brief multisensory exposure, especially when stimulus congruency, timing, or task relevance are not optimal (e.g., Molholm et al., 2006; Stevenson et al., 2012). These mixed findings suggest that the influence of prior multisensory experience may be contingent on contextual and stimulus-driven factors, including temporal alignment, semantic congruency, task design, individual differences and attentional demands. Overall, these results highlight the role of audio-visual integration in

shaping object representations and suggest that prior multisensory experience might modulate early perceptual stages during later unisensory recognition.

Evidence for multisensory benefits on object perception suggests that the brain is inherently multisensory (Ghazanfar & Schroeder, 2006). However, while the role of multisensory inputs in object perception and recognition is well-established, the processes governing multisensory object categorisation remain not fully understood. In particular, it is unclear how information from multiple senses contributes to the formation of object categories. On one hand, multisensory object representations may support more efficient identification and differentiation among objects, as they provide redundant and complementary information across modalities (e.g.; Thelen et al., 2012). For example, combining visual and haptic cues can enhance perceptual precision and improve object recognition, especially under sensory uncertainty. On the other hand, some researchers have argued that object categories are abstract and amodal in nature, existing independently of specific sensory modalities (e.g., Pylyshyn, 1980; Mahon & Caramazza, 2008; Binder, 2016; Leshinskaya & Caramazza, 2016). According to this view, sensory information is transformed into modality-independent conceptual representations, supporting categorisation regardless of the input channel. A better understanding of how multiple senses contribute to the categorical perception of objects might reveal the processes underlying the formation of object categories in memory.

2.3 Outline of the Experiments

In Chapter 2 we aim to investigate whether multisensory information influences the categorisation performance of familiar objects. This investigation is conducted across three experiments. In Experiment 1A, we assessed the representativeness of auditory stimuli for their corresponding visual object categories. Participants were presented with object sounds (i.e., the sound of a bell) and asked to match the sound to the corresponding object category (i.e., ‘Bell’). They first had to match the sound to the relative basic-level object category label (i.e. sound-

object category task), and then further identify the specific exemplar related to the sound (i.e. sound-object exemplar task). This assessment is important to establish that the sounds adopted in these experiments can reliably inform category membership, as auditory cues are often not integrated unless they are explicitly task-relevant or semantic informative (Bizley et al., 2016; Chen & Spence, 2010). Moreover, in Experiment 1B we investigated whether prior multisensory audio-visual exposure to objects category affected participants' subsequent visual categorical perception of familiar objects. Using an adapted version of the paradigm developed by Murray et al. (2004, 2005, 2013), participants first experienced objects with semantically congruent visual and auditory cues and were then assessed on their ability to categorise and discriminate the objects, in a visual only categorical perception task. By presenting audio-visual categories prior to test, Experiment 1B evaluates whether multisensory exposure strengthens category representations compared to visual-only exposure. Finally, in Experiment 1C, we investigated the perceptual similarity of the morphed object continua adopted in Experiment 1B, to verify that perceptual similarity judgments increased monotonically with morphing level, and that intermediate stimuli were judged more similar to both category endpoints than to unrelated categories.

2.4 General Stimuli

We designed object shapes (3D-designed) with SketchUp Pro (www.sketchup.com). Initially, the outline shapes for each object were created and standardised in size (i.e., maximum height of 100 mm and width of 70 mm), the outlines were then rotated along their vertical axes to create their corresponding 3D object. The object stimuli used in these experiments are four basic-level category objects with two sub-ordinate level objects from each basic level as follows: Bell (Hand, Church), Bottle (Wine, Coke), Glass (Wine, Beer), Vases (Urn, Single Stem), see Figure 2 for examples of the stimuli. These objects were then morphed to create a series of object-shape continua using the 3D modelling software Blender v3.3.0 (2022; www.blender.org), following a

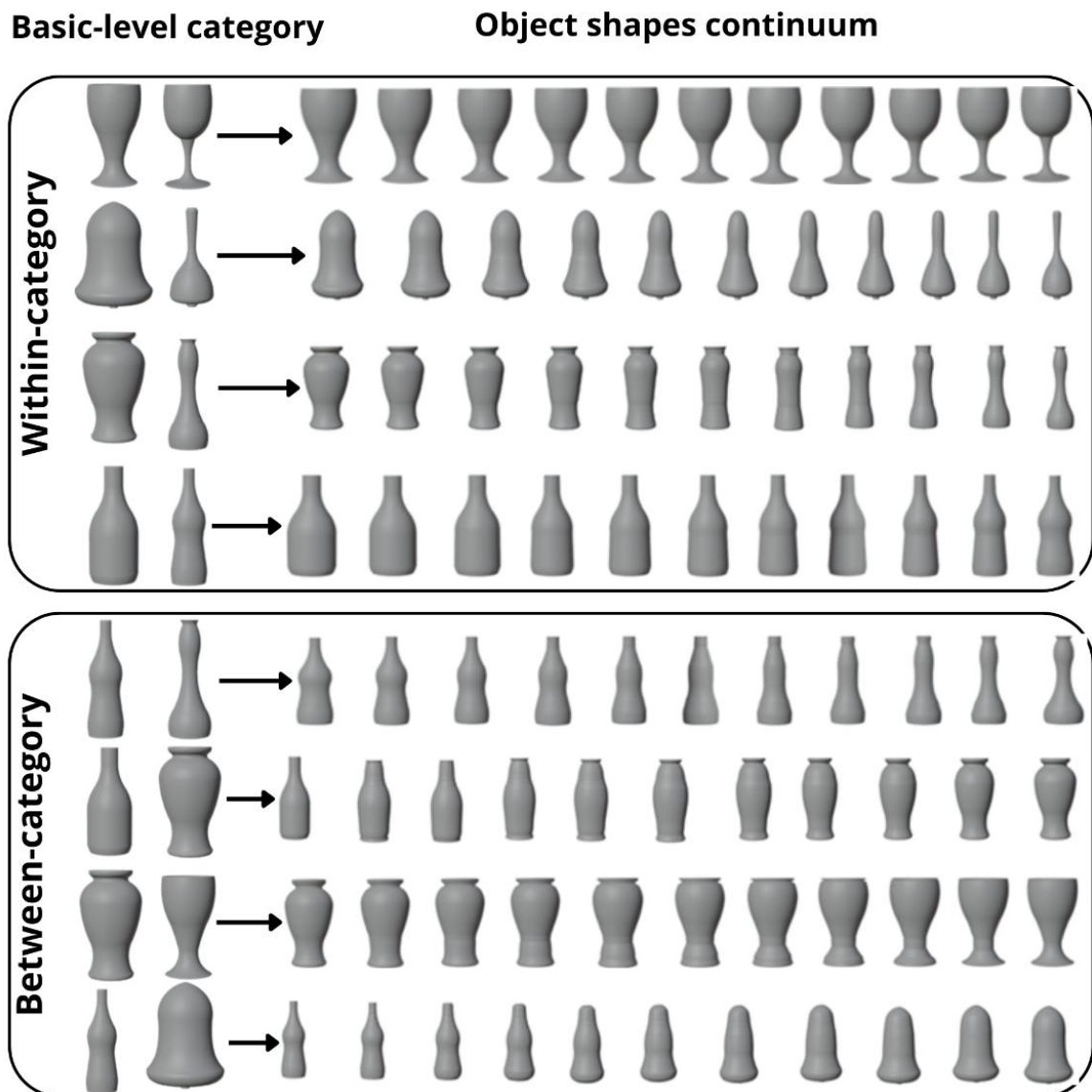
procedure previously described in other study (Newell & Bühlhoff, 2002). Specifically, for each pair of objects, we created a morphed continuum of eleven images, including the two original object images, one at each extreme end of each continuum and nine morphed object images. Within Blender workspace, we applied the Shrinkwrap Modifier (in Modifier Properties) to one of the two objects in each pair and assigned the Shape Key to the other object (range 0 to 1). Every object in each pair was created by keeping constant a 0.1 step distance from each other. The objects were extracted from Blender while keeping both the camera viewpoint and the lighting condition constant. The main camera location was kept stable at the central point of the 3D space (location [0 m, -0.4 m, 0.1 m], rotation [8°, 0.07°]). The light source (point, radius 0.1 m, 1000 W) was kept constant at the central point of the 3D space (location [-0.4 m, -0.4 m, 0.5 m], rotation [61°, 21°]). The render engine Workbench, with Studio Lighting, Colour Material (0132, light grey), and Specular Lighting, was applied to each object. All the images were extracted with a resolution of 1080 x 1080 px, at 100% output scale (i.e., meaning full resolution was retained without downscaling or compression during export), and were presented in a canonical view, i.e., a familiar view of the objects in which most features can be viewed (Cutzu & Edelman, 1994), against a transparent background. We created morphed 3D objects using objects from either within the same basic-level category (e.g. wine to beer glass) or between basic level categories (e.g. wine glass to cola bottle) yielding to a total of twenty-eight objects continua, four within-category and 26 between-category morphs (see Figure 2 for an example of the stimuli).

The list of morphed continua adopted in this Chapter is as follows: Bell (Church – Hand), Bottle (Wine – Coke), Glass (Wine – Beer), Vase (Single stem – Urn), Church Bell - Wine Bottle, Church Bell - Coke Bottle, Hand Bell - Wine Bottle, Hand Bell - Coke Bottle, Church Bell - Wine Glass, Church Bell - Beer Glass, Hand Bell - Wine Glass, Hand Bell - Beer Glass, Church Bell - Single stem Vase, Church Bell - Urn Vase, Hand Bell - Single stem Vase, Hand Bell - Urn Vase, Wine Bottle - Wine Glass, Wine Bottle - Beer Glass, Coke Bottle - Wine Glass, Coke Bottle - Beer Glass, Wine Bottle - Single stem Vase, Wine Bottle - Urn Vase, Coke Bottle - Single stem Vase, Coke Bottle -

Urn Vase, Wine Glass - Single stem Vase, Wine Glass - Urn Vase, Beer Glass - Single stem Vase, Beer Glass - Urn Vase.

Figure 2

An illustration of examples of familiar objects adopted in Experiment A, 1B and 1C, grouped by basic level category and the object shapes continua.



Note: Within-category objects are the Bell (hand, church), Bottle (wine, coke), Glass (wine, beer), and Vase (urn, single stem) pairs, while the between-category objects illustrated here are coke Bottle - Single stem Vase, urn Vase – wine Bottle, urn Vase – beer Glass and coke Bottle – church Bell pairs.

In Experiment 1B, the distractors stimuli used during the old-new recognition task (i.e., Stationery (Pen, Highlighter), Phone (Smart, Home), Clock (Wall, Wrist)) were downloaded from publicly available online 3D model repositories (e.g., TurboSquid, Sketchfab), then rendered into 3D-images using a standardised virtual camera setup to ensure consistent lighting, perspective, and resolution across stimuli. The stimuli adopted in the irrelevant condition are alphabet letters images presented either in capital or italic format (e.g., A, B, C, D; a, b, c, d), and the letter stimuli (A–J) were extracted using a monospaced font (OCR-A), designed for optical character recognition and helping high visual distinctiveness (Ginger & Spiekermann, 2003). The images of letters were displayed in black on a transparent background, using a sans-serif font (Arial), and they were presented at a standardised at a size of (0.5, 0.5) in PsychoPy's norm units, corresponding to 50% of the display window in both height and width.

Sound clips were downloaded from the web from a number of available sound databases (e.g., Free Sounds Library, freeSFX, Storyblocks). A total of 16 sound stimuli were used in these Experiments. Each sound matched an exemplar from the visual objects (i.e., hand bell, church bell, beer glass, wine glass, single stem vase, urn vase, wine bottle, coke bottle). The speech sound of a letter was also paired with its corresponding visual letter for use in the experiment. The sound clips were edited using Audacity® (v.03, <https://audacityteam.org/>), and all clips were presented for a 500 ms and at a rate of 44,100 Hz, referring to the audio waveform was digitised at 44,100 samples per second to ensure high-quality playback across devices.

Experiment 1A, 1B and 1C were developed in PsychoPy (Pierce et al., 2019, 2022, v2023.2.3) and run online through Pavlovia (<https://pavlovia.org/>) on desktop devices.

2.5 Experiment 1A

In Experiment 1A, we assessed the representativeness of the sounds created to represent the semantically congruent audio-visual modality, since auditory information is usually only integrated when it is providing congruent semantic information or task-relevant (Chen & Spence,

2010; Bizley et al., 2016). To this end, we tested participants' ability to correctly pair sounds with their corresponding object categories in two auditory identification tasks. Participants first heard a sound and were asked to identify the object category it belonged to (e.g., glass), and then to indicate which specific object exemplar within that category the sound was most likely associated with (e.g., wine glass vs. beer glass). Stimuli in this task were presented either as images or as written object names. For example, participants could hear a sound of clinking glasses, and then first select to which object the sound would belong, a glass in this example. Subsequently, they were tasked with selecting whether the sound was related to a beer glass or a wine glass. We expected high accuracy in matching sounds to their corresponding object categories, particularly in the first-level category decision (i.e., Sound-object category matching task), as the auditory stimuli were designed to be semantically congruent with familiar everyday objects. This would confirm the effectiveness of the audio recordings as category-level cues. For the second-level identification (i.e., Sound-object exemplar matching task), we anticipated more variability, reflecting the greater perceptual and semantic ambiguity at the exemplar level. The inclusion of both image- and name-based response formats allowed us to explore whether visual familiarity influenced exemplar-level identification. Successful performance in this task aimed to validate the use of these sounds as meaningful auditory cues in the multisensory learning phase of Experiment 1B.

2.5.1 Method

Participants

We recruited thirty-two adult participants (mean age = 41.5 years, SD = 12.7; 50% female) for Experiment 1A using Prolific (<https://www.prolific.co/>), the majority of whom completed at least second-level education (99.8%; 82.3% completed third level education). Our inclusion criteria were as follows: fluency in English, normal or corrected-to-normal vision and no hearing impairments. All the participants were naïve to the purpose of the study and received compensation to take part in the experiment (at a rate of £9.00 per hour). The study was approved

by the Trinity College Dublin, School of Psychology Research Ethics Committee (approval number SPREC102020-50) and complied with EU data protection legislation (GDPR).

Stimuli

The stimuli used in Experiment 1A are the extremes of each object continuums, or ‘basic-level category’ objects, and the correspondent auditory stimulus, for a total of eight object-sounds pairings, see General Stimuli and Figure 2.

Design

Experiment 1A was based on a mixed design, with *object category* (4 levels: bell, glass, vase, bottle) as a within-subject factor and *object version* (2 levels: images or labels) as a between-subject factor. Participants performed a two-level auditory matching task: after a training phase to familiarise with the task, participants heard a sound and were asked to first match the sound with the basic-level object category label (‘Bell,’ ‘Bottle,’ ‘Vase,’ ‘Glass’; i.e. sound-object category task). Once the object category was selected, participants had to further identify the specific exemplar to which the sound was related (i.e. sound-object exemplar task). In the sound-object exemplar matching task, exemplars were either presented as visual stimuli or labels. The two tasks were presented in blocks, and the trials within each block were fully randomized across participants. The main dependent variable in this experiment was identification accuracy, and we also collected reaction times (in milliseconds) for completeness. The study had a total of 16 (8 x 2, n of sounds x n of repetitions) experimental trials. The whole experiment took approximately 10 minutes to complete.

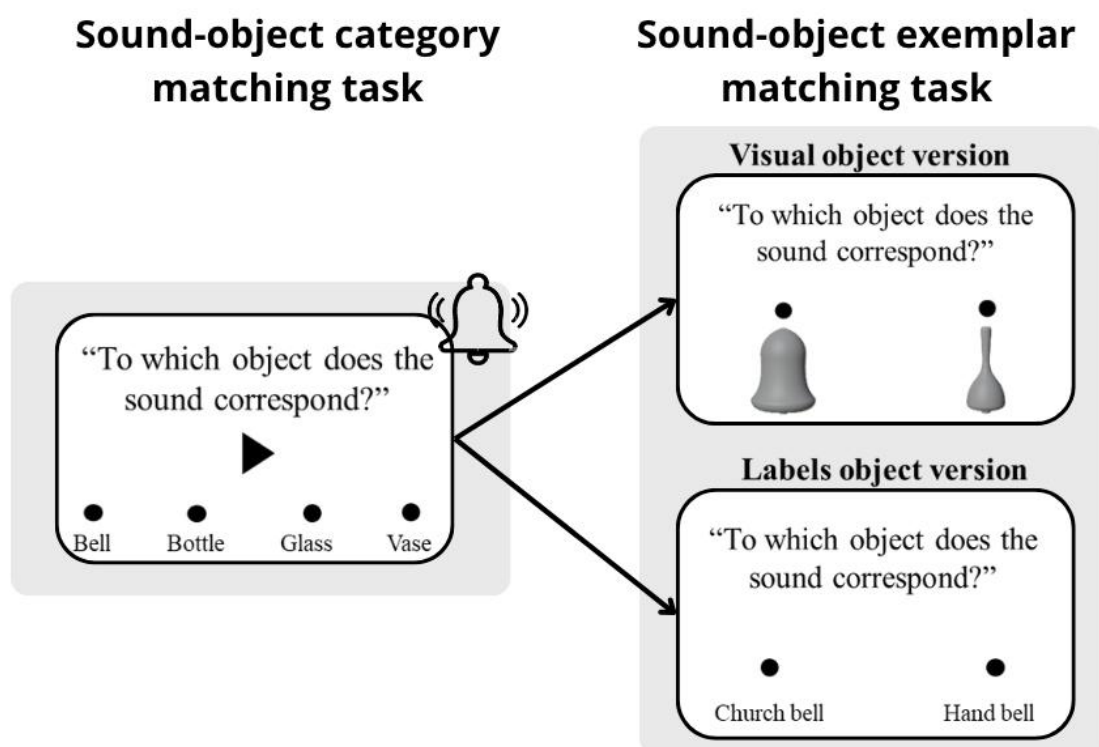
Procedure

Prior to the experiment, all participants provided informed consent to proceed. This was then followed by a basic questionnaire on demographic information. The task instructions explained that their task was to identify which object the target sound belonged to, and then which specific exemplar of that object the sound belonged to, as quickly and as accurately as possible. The experiment then began with a short training, followed by the main task. A question

asking, ‘To which object does the sound belong?’ and a button to play the sound appeared at the top and bottom of the screen, respectively. The sounds played for 500ms, max repetition of two times, after which participants had to select one of the four options: ‘Bell,’ ‘Bottle,’ ‘Glass,’ or ‘Vase’ as the correct match to the sound. Once the response had been recorded, participants were then asked, within the same trial, to select the specific exemplar of the selected object category, either presented through labels or images, that matched the sound. Participants made their response using a mouse click. For example, if a participant had initially selected ‘Bell’, they would then need to further indicate whether the sound matched a hand bell or church bell, as presented either as words (labels) or images of 3D objects. The pair of objects or word labels were presented on the bottom left and right of the screen, and they remained on screen until a response was made. Each trial was followed by a 500ms inter-trial interval. See Figure 3 for a schematic illustration of an example trial.

Figure 3

Example of an experimental trial in Experiment 1A.



Note: On the left is portrayed the sound-object category matching task, followed by the two alternative versions (visual and labels objects) of the sound-object exemplar matching task.

Analysis

First, we assessed whether participants correctly matched the auditory stimuli to the corresponding object class in the sound-object category matching task. In this task, each object category was tested across four trials, yielding accuracy scores of 0 (never correct), 0.25 (correct once), 0.5 (twice), 0.75 (three times), or 1 (always correct). Scores of 0 and 0.25 were categorised as ‘Below chance’, 0.5 as ‘At chance’, and 0.75 or 1 as ‘Above chance’. These categorical accuracy levels were analysed using a generalised logistic mixed-effects regression (glmer) model, with object category as fixed effect. For the analysis of sound-object exemplar matching task performance, we included only trials in which participants correctly identified the sound-object category in the previous task, thereby ensuring that responses reflected accurate sound-object associations. Thus, in this task, each stimulus was tested twice, yielding accuracy scores of 0, 0.5, or 1. Scores below 0.5 were coded as ‘Below chance’, 0.5 as ‘At chance’, and 1 as ‘Above chance’. A separate generalised logistic mixed-effects regression (glmer) model assessed whether the object version predicted performance, with fixed effects object category, object version, and their interaction. All models adopted a logit link and the ‘bobyqa’ optimiser, and controlled for age, sex, and reaction times as covariates. All analyses were conducted in R using RStudio (R Core Team, 2021). Models were fit with the lme4 package (Bates et al., 2014), and post-hoc comparisons were performed using the emmeans package (Lenth Russell et al., 2022), applying Bonferroni correction for multiple comparisons. The Experiment 1A project overview and analysis scripts are available on the Open Science Framework: https://osf.io/uec6z/?view_only=1Cc9ea6c0667403e9cfd7bb8332b1B7f.

2.5.2 Result

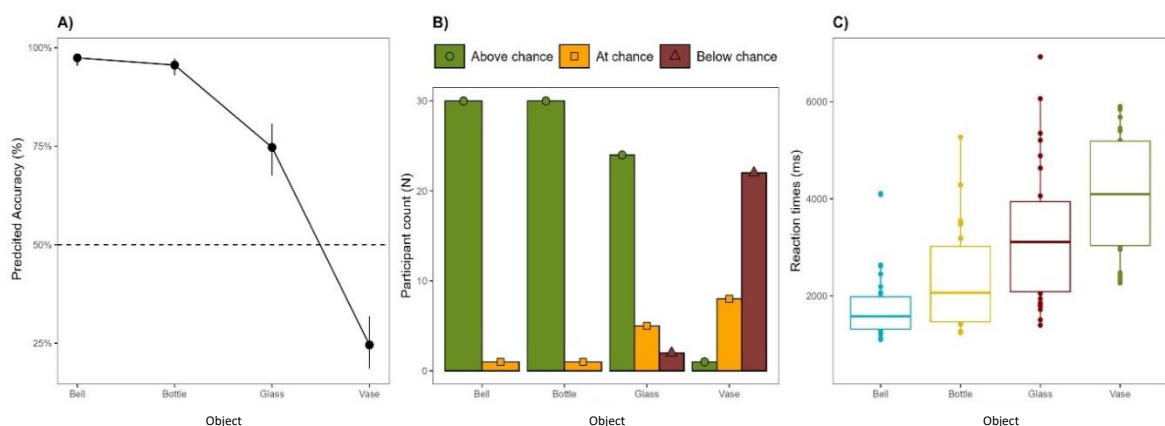
Sound-object category matching data. Mean performance is reported in Table 1.

Table 1*Sound-object category matching task performance for object category in Experiment 1A.*

Object category	Accuracy		Reaction times (ms)	
	Mean	SD	Mean	SD
Bell	0.97	0.16	1804	984
Bottle	0.95	0.23	2365	1522
Glass	0.71	0.46	3316	2150
Vas	0.25	0.43	4220	2644

Note: SD stands for standard deviation.

A comparison between models, with and without object category as a factor, was performed, and a likelihood ratio test suggested that object category significantly contributed to the model predicting the probability of a correct matching response, $\chi^2(3) = 699.4$, $p < .001$, as shown in Figure 4A. The sound stimuli for Vases were 4.76 times less likely to be correctly matched to its object compared to sounds of Bells (95% CI [0.12, 0.80]). Glass sounds were 1.55 times less likely to be correctly matched to their objects than Bell sounds, while Bottle sounds were 0.44 times less likely to be correctly matched to their object compared to Bell sounds. The overall performance in the sound-object category matching task is shown in Figure 4B and 4C.

Figure 4*Overall performance for the sound-object exemplar matching task in Experiment 1A.*

Note: The plot shows the summary of findings of mixed-effects models predicting probability of correct identification response in the sound-object category matching task across for each

object category (panel A), the dashed horizontal line demarcates chance level (50%). The number of participants who scored above chance, at chance and below chance for each object (panel B). Performance for each object category in the different versions for reaction times (panel C).

Sound-object exemplar matching task. Mean performance is reported in Table 2.

Table 2

Sound-object exemplar matching task performance for object exemplars in Experiment 1A.

Object exemplar	Accuracy		Reaction times (ms)	
	Mean	SD	Mean	SD
Beer Glass	0.44	0.50	2762	1258
Church Bell	0.44	0.50	2033	883
Coke Bottle	0.48	0.50	2089	813
Hand Bell	0.53	0.50	1967	716
Single stem Vase	0.41	0.50	2332	934
Urn Vase	0.5	0.50	2679	1224
Wine Bottle	0.59	0.50	2434	1266
Wine Glass	0.45	0.50	2576.	1198

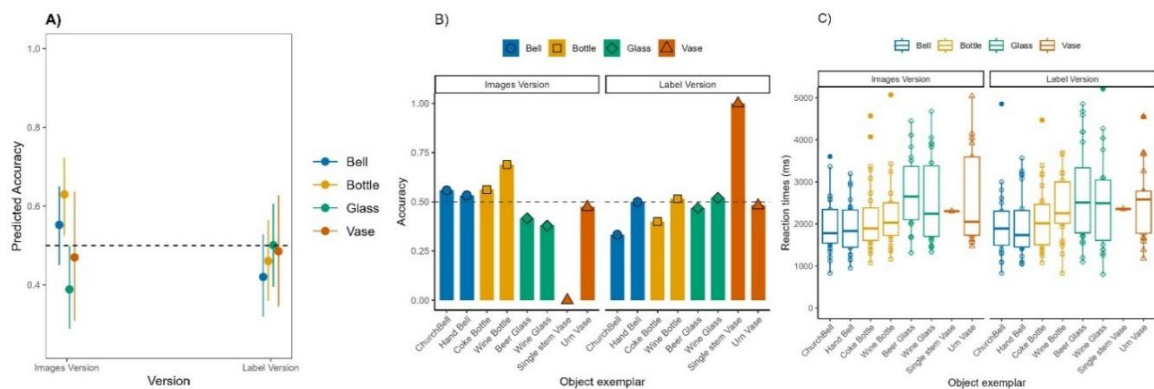
Note: SD stands for standard deviation.

For the analysis of sound-object exemplar matching task performance, we included only trials in which participants correctly identified the sound-object category in the previous task, thereby ensuring that responses reflected accurate sound-object associations. A comparison between models, with and without the object category by object version interaction, was performed. A likelihood ratio test indicated that the interaction significantly improved the model predicting the probability of correct identification, $\chi^2(3) = 10.39$, $p = 0.012$, showing better fit for the interaction model compared to the main-effects-only model. Predicted probabilities revealed that accuracy varied across object category and between object versions, see Figure 5A. Regarding the Images-Version, predicted accuracy ranged from 39% (95% CI [29%, 50%]) for Glass to 63% (95% CI [53%, 72%]) for Bottle sounds, whereas for the Label-Version, accuracy was generally lower for Bell

(42%, 95% CI [32%, 53%]) and Bottle (46%, 95% CI [36%, 56%]) but comparatively higher for Glass (50%, 95% CI [40%, 61%]) and Vase (49%, 95% CI [35%, 63%]) sounds. Odds ratios derived from fixed effects showed that Glass and Vase sounds had significantly lower odds of correct identification relative to Bell sounds. Specifically, Glass sounds were approximately 2.7 times less likely to be correctly identified than Bell sounds overall (OR $\approx$ 0.52; 95% CI not reported here) but exhibited a double increase in accuracy under the Label Version condition (interaction OR $\approx$ 2.7). Similarly, Vase sounds showed a fivefold reduction in odds compared to Bell sounds (OR $\approx$ 0.72) but also benefited disproportionately from the Label-Version (interaction OR $\approx$ 1.8). These results demonstrate a complex modulation of identification accuracy by object category and object version, with interactions suggesting differential effects of label versus image-based cues. The overall performance in the sound-object exemplar matching task is reported in Figure 5B and 5C.

Figure 5

Overall performance for the sound-object exemplar matching task in Experiment 1A.



Note: The plot shows the summary of findings of mixed-effects models predicting probability of correct identification response in the sound-object exemplar matching task across different object versions and for each object category (panel A). The dashed horizontal line demarcates chance level (50%). Performance for each object exemplar in the different versions for accuracy (panel B) and reaction times (panel C).

2.5.3 Discussion

The results of Experiment 1A suggested that overall sounds were correctly matched at above-chance levels. Specifically, during the sound-object category matching task, participants were able to correctly match sounds related to the Bell at above-chance levels. In contrast, sounds related to the Vase were matched significantly less accurately compared to Bell sounds, suggesting that Vase sounds may not sufficiently represent either the basic-level category or its exemplars. Notably, sounds related to Glass were twice as likely to be correctly matched to their object in the label condition compared to the image condition, suggesting that participants were better able to associate sounds with the abstract category label than with the 3D visual object representation. This discrepancy may reflect individual variability in participants' mental representations of the Glass category, such that the Glass sounds did not adequately correspond to their personal prototypes of the object. Alternatively, the sound of a glass, such as clinking, may activate a dynamic representation of the object that the label captures, whereas the static images used to represent the Glass category lack this dynamic quality. Overall, Bell sounds were significantly representative of their basic-level category but did not support within-category exemplar identification. The difference in representativeness between Bell and Vase sounds may be attributed to the nature of the sounds: Bell sounds are directly related to the object's action (e.g., ringing of a church bell), whereas Vase sounds may primarily reflect the material properties of the object, aligning more with a representation of the material itself rather than the object as a whole. Finally, during the object exemplar matching task, sounds were overall more likely to be correctly matched with the object image than with the label version, suggesting that images generally provided a more accurate representation of the objects. However, this effect varied by object category, with Glass and Vase sounds benefiting relatively more from the label version compared to other object categories.

2.6 Experiment 1B

Experiment 1B aimed to assess whether exposure to bimodal (audio-visual) object categories influences the categorisation process of corresponding visual stimuli. If multisensory information of objects is stored as multisensory categories in memory, exposure to bimodal (audio-visual) semantically congruent categories should significantly enhance subsequent unisensory (visual only) categorical perception. Specifically, bimodal learning should significantly shape the categorisation function by inducing better discrimination performance and sharpening the category boundary, compared to the unisensory condition, if this assumption holds valid. To test this, In Experiment 1B we assessed participants' unimodal categorical perception (CP) performance following a pre-exposure phase in which participants were presented with target objects across different modality conditions. Specifically, participants first completed a Learning phase in which they were exposed to target objects under four auditory conditions: semantically congruent sounds (e.g., an image of a wine glass paired with the sound of clinking glass; AV_c), incongruent sounds (e.g., the same image paired with a flushing water sound; AV_{inc}), no sound (unisensory visual exposure; V), and an irrelevant audiovisual condition (AV_{irr}) in which participants were presented with images and sounds of alphabet letters unrelated to the object stimuli. This design allowed us to assess whether the semantic relationship between modalities, rather than audiovisual pairing, influenced later visual categorisation. The purpose of the AV_{irr} condition was to control for general multisensory exposure effects and to isolate the role of meaningful cross modal associations. Following this, participants completed an old-new recognition task to assess memory for the objects presented during exposure. After it there was the Categorical Perception phase, in which participants were presented with visual-only (no sound) stimuli sampled from morphed continua between the previously encountered target objects. They completed both a discrimination task (judging whether a morph resembled prototype A or B) and a categorisation task (assigning it to Category 1 or 2 based on shape).

We predicted that congruent audio-visual exposure condition would facilitate stronger and more integrated object representations, leading to enhanced categorical perception. Specifically, we predicted greater discrimination accuracy and faster response times for stimuli near the category boundary, where perceptual uncertainty is highest, for objects previously exposed in the congruent condition compared to those in the incongruent condition. This boundary advantage was expected to reflect a sharpening of category representations driven by semantically aligned multisensory input, as reflected in sharper category boundaries during the visual-only categorisation task. In contrast, incongruent exposure condition was expected to disrupt perceptual encoding due to semantic conflict between modalities, resulting in weaker categorical perception. The no sound condition served as a baseline for unisensory learning, allowing us to assess whether audiovisual pairing confers a multisensory advantage. The structure of the paradigm thus enabled us to examine whether the semantic congruency of audiovisual information during exposure modulates subsequent unisensory categorisation performance.

The analysis in Experiment 1B were designed to address three main research questions. First, we asked whether visual stimuli are categorically perceived when learned through unisensory (visual-only) exposure. Categorical perception is typically demonstrated when discrimination performance peaks for stimuli at the category boundary—defined as the point along the stimulus continuum at which responses are at 50% for each category—while stimuli at the extremes elicit lower discrimination sensitivity (Harnad, 1987; Goldstone, 1994b). Accordingly, we expected more accurate and faster discrimination for boundary stimuli compared to those at either end of the continuum. Second, we examined whether the position of the category boundary varies depending on the learning modality (i.e., congruent, incongruent, unisensory, or irrelevant). If learning with semantically congruent audiovisual information systematically shifts the perceived boundary, this would indicate that multisensory exposure not only strengthens category representations but may also alter their perceptual structure. We also explored whether any such shift is object-independent, suggesting a generalisable effect of semantic audiovisual congruency.

Third, we investigated whether exposure to congruent audiovisual pairings leads to improved overall categorisation and discrimination performance, as reflected in accuracy and reaction times (RTs), compared to other learning conditions. Performance following unisensory visual learning was used as a baseline against which to assess the benefit of meaningful multisensory associations. To further characterise the nature of categorical perception, we examined the steepness of the categorisation function, or *slope*, across learning conditions. A steep slope indicates a sharp perceptual transition between categories, suggesting efficient category formation, whereas a shallow slope implies a more gradual perceptual shift. Exclusion criteria were defined prior to testing, and these were based on data obtained from a preliminary pilot study.

2.6.1 Method

Participants

All participants were recruited using Prolific (<https://www.prolific.co/>), based on the following inclusion criteria: fluency in English, normal or corrected-to-normal vision and no hearing impairments. Only participants who had access to a desktop computer could take part in the study. An a priori power analysis for each of the groups was performed in PANGEA (Westfall, 2016; v0.2) for a mixed-subjects design. Based both on the power analysis and on previous literature (Bülthoff & Newell, 2002, 2004; Calder et al., 1996), the advised minimum sample size is for 80% power to detect the effects of interest with size (Cohen's *d*) of 0.3, was 96 participants, with *object-category* as factors (2 levels each), and *learning conditions* (4 levels). We initially recruited 108 participants of whom 101 (94%) successfully reached the learning criteria. In total, 101 participants (mean age = 33.1 in years, SD = 11.4, 52.5% female) completed the experiment online. All participants were naïve to the purpose of the study and were compensated at a rate of £9.00 per hour. The study was approved by the Trinity College Dublin, School of Psychology

Research Ethics Committee (approval number SPREC102020-50) and complied with GDPR data protection legislation.

Stimuli

The stimuli adopted in Experiment 1B are described in the General Stimuli section. In the exposure task of the Learning phase, the stimuli presented are images-sound pairings: we paired the extremes of each object continuums, or ‘basic-level category’ objects (see Figure 2), with either semantically congruent, incongruent, or irrelevant sounds (i.e., letters), or presented with no audio (visual presentation only). The stimuli presented in the visual Categorical Perception phase, are the object shape continua described in the General Stimuli section (see Figure 2). Twenty-eight object continua were adopted in Experiment 1B, including four within-category morphs and twenty-four between-category morphs, calculated as follows: 4 within-category + 24 -between-category (6 between-category comparisons $\times$ 2 exemplars per category $\times$ 2 exemplar per category).

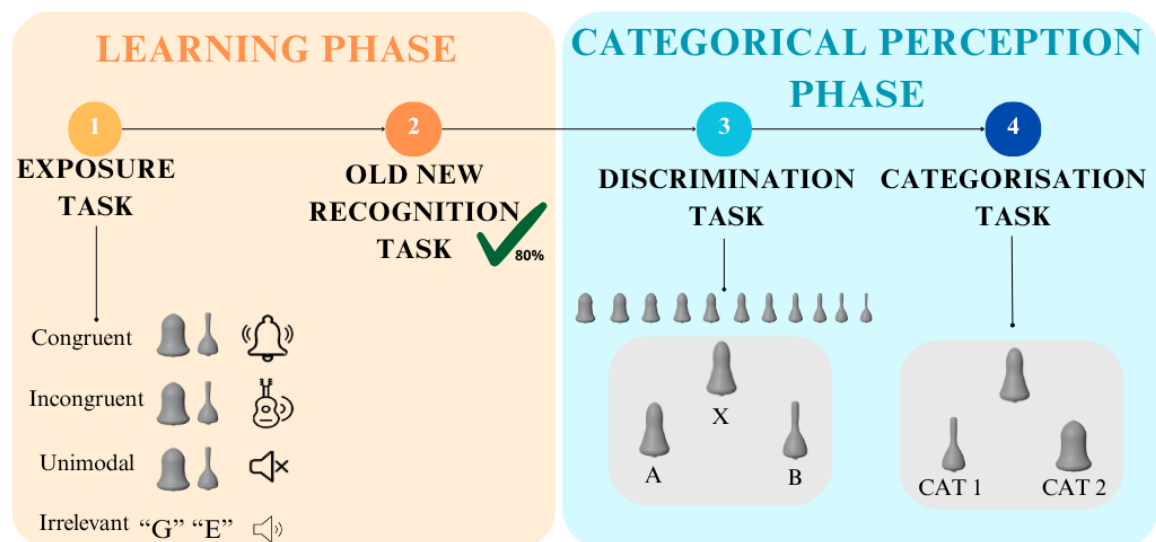
Design

There were two main phases to the experiment: a Learning phase and a (visual only) Categorical Perception phase. Each participant performed the Learning phase followed by the Categorical Perception phase, with a self-timed break between the two, see Figure 6 for a visual illustration of the paradigm. The experiment was based on a mixed design, with learning conditions as a between-subject factor (4 levels: audio-visual semantic congruent, audio-visual semantic incongruent, audio-visual irrelevant, visual only) and object category as a within-subject factor (2 levels: within and between). During the exposure session, each pair of familiar objects from the different basic level categories were presented with a sound or visually only. The paired sounds were either semantically congruent, AV_c (e.g. image and sound of a church bell), semantically incongruent, AV_{inc} (image of church bell and sound of wine pouring), or no sound was presented (V). To isolate the potential effect of audio-visual combinations on object perception,

an additional control condition was designed, consisting of the images and sounds of letters (AV_{irr}). This condition served to control for non-specific effects of audio-visual exposure and task repetition, ensuring that any observed effects in the experimental conditions were due to semantic congruency, and not to the presence of multisensory input. This approach is consistent with prior work that has used semantically incongruent or neutral audiovisual pairings to distinguish meaning-based multisensory integration from low-level crossmodal processing (e.g., Chen & Spence, 2010; Laurienti et al., 2004).

Figure 6

Overview of the experimental paradigm in Experiment 1B.



Note: The study comprised a Learning phase and a Categorical Perception phase. In the Learning phase, participants first completed an exposure task, followed by an old-new recognition task assessing memory for previously seen objects. In the Categorical Perception phase, participants performed a discrimination task involving morphed stimuli along a continuum between two shape categories, and a categorisation task requiring binary classification of test items based on prior learning. All stimuli in this phase were presented visually, without sound.

Learning Phase. The Learning Phase was based on a single factor, within subject's design with four learning levels (AV_c, AV_{inc}, V and AV_{irr}), and comprised both an exposure and test task.

During the initial exposure task, participants were asked to learn the specific audio-visual combination of each object, and each image object was presented twice. Following the exposure, participants then performed an old-new recognition task (Lehmann & Murray, 2005) to assess whether they learned the specific objects-sounds combination. Specifically in exposure task, participants perceived the following stimuli across different learning levels (AV_c , AV_{inc} , V and AV_{irr}): Bell (Church–Hand), Bottle (Wine–Coke), Glass (Wine–Beer), Vase (Single stem–Urn), and the letter pairs A–a, B–b, C–c, and D–d. During the subsequent test task, the same stimuli served as the ‘old’ or target stimuli, while novel items were introduced as ‘new’ distractors, including Lamp (Bedside–Desk), Stationery (Pen–Highlighter), Phone (Mobile–House), Clock (Wall–Watch), and the letter pairs E–e, F–f, G–g, and H–h. To control for item-specific effects, object pairings were fully counterbalanced across twelve participant groups. Each group was exposed to a different combination of four object pairs (e.g., Bell, Bottle, Glass, Vase) during the Learning phase, with each pair drawn from a continuum (e.g., Hand–Church, Wine–Coke). The order of stimulus presentation within each task was randomised for each participant. Refer to Supplementary Materials - Table S2 for details of the group-wise counterbalancing scheme of the learning conditions within participants in each group. A threshold of 80% accuracy was required to continue to the Categorical Perception phase, which if not met the Learning Phase was repeated up to a total of three times. If the participant failed to reach the threshold accuracy criterion after three repetitions, the study was concluded. In the Learning Phase, there were a 16 (8×2 , n of objects $\times$ n of repetitions) experimental trials in the exposure task and 32 (16×2 , n of objects $\times$ n of repetitions) experimental trials in the test task.

Categorical Perception Phase. During the Categorical Perception phase, participants were tested only with images of objects (i.e. no sounds) presented during the Learning phase. The structure of the visual-only Categorical Perception phase was the same for all participants: a discrimination task was followed by a categorisation task, with self-timed breaks between the tasks. The categorisation task aimed to determine the subjective category boundary for each

object pair; thus, this task was performed after the discrimination task to avoid any bias during the latter. During this phase, participants were asked to categorise individual images from the morphed continua created. Both in the discrimination and categorisation tasks, objects pairs were presented in separate test blocks. To control for item-specific effects, object pairings were fully counterbalanced across twelve participant groups. Participants were randomly assigned to different pairs of objects for testing, and the assignment of object pairs to within-category and between-category conditions was counterbalanced. During the experiment, each participant was tested on four pairs of objects: 2 within-object and 2 between-object classes, with each pair representing one of the learning conditions (i.e. AV_c , AV_{inc} , V and AV_{irr}). The participants in each group perceived the same objects in the Learning and Categorical Perception phases. A total of twelve different combinations of testing objects was created, with participants assigned to one (refer to Supplementary Materials - Table S2 for a detailed description of the objects presented in each group). In the discrimination task there were 144 ($4 \times 9 \times 4$, n of object pairs x morphed images x XAB counterbalancing) experimental trials and the categorisation task had 72 ($4 \times 9 \times 2$, n of object pairs x n of morphed images x n of repetitions) experimental trials. The main dependent variable in this experiment was accuracy although we also collected reaction times (RTs) for completeness. Each participants took approximately 15 minutes to complete the experiment.

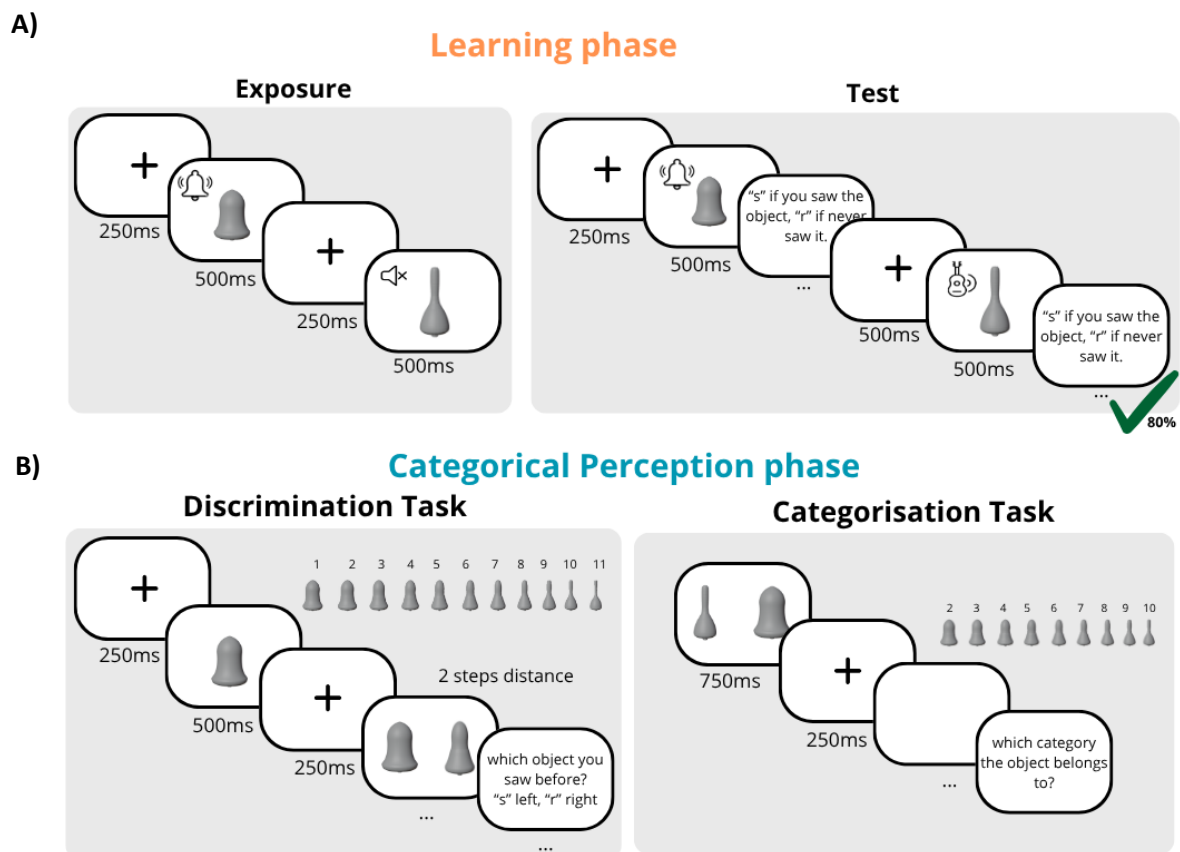
Procedure

Participants were explicitly instructed to complete the study in a quiet room and to wear headphones throughout the session to ensure optimal delivery of the auditory stimuli, and they provided responses using their keyboard. Participants were first asked to read the study information sheet, provide informed consent and provide some basic demographic information. The experiment started with a sound check, asking participants to adjust the volume level of their speaker so that they could clearly and comfortably hear the sound. The Learning Phase was

followed by the Categorical Perception phase, with a self-timed break between the two phases, see Figure 7.

Figure 7

An illustration of a trial sequence from the audio-visual congruent learning condition in Experiment 1B.



Note: The image shows the Learning phase (panel A), both exposure and test tasks (bimodal example), and the visual-only Categorical Perception phase (panel B), both discrimination and categorisation tasks.

During the Learning phase, participants were initially shown the category extremes of four target objects (e.g., exposure task), after which they were tested (e.g., test task) on their ability to recognise the objects. During the exposure task, participants were instructed to learn the combination of audio-visual objects presented, and no responses was required. A fixation cross preceded the object stimuli in all trials for 250 ms, each stimulus was presented on the screen for 500 ms, followed by a 250 ms inter-trial interval. Following exposure, participants saw the same objects paired with congruent or incongruent sounds, with no audio or in the irrelevant condition

(i.e., female voices presented with capital letters, and male voices with letters in italics). In the test task, participants were presented both with target objects previously seen in the exposure task, and distractor objects. For each trial, participants performed a 2-alternative forced-choice (2-AFC) task in which they had to decide whether the object was already presented, or it was novel. Participants responded pressing the 'S' key if they had seen the object before and the 'L' key if the object was new. The feedback *'Too slow! Respond faster in the next trial'* was provided if responses were provided after two seconds from the start of stimulus onset. A threshold of 80% accuracy in recognising the target objects was required to continue to the Categorical Perception Phase.

Once the Learning phase was completed, participants performed the visual-only Categorical Perception phase, consisting of a discrimination task followed by a categorisation task, with self-timed breaks between the tasks and no feedback provided. The discrimination task was based on an XAB design: in each trial, the first image (X) was presented, followed by two other object images (A and B) which were simultaneously presented to the left and right of the centre of the screen. Stimuli A and B always differed by two steps (e.g., image 5 and image 7). Stimulus X was always identical to either A or B. The order of stimuli was counterbalanced, resulting in four different orders of presentation for any two stimuli (AAB, ABA, BBA, BAB). In all trials, a fixation cross (250 ms) preceded the presentation of stimulus X. The first object (stimulus X) was presented in the centre of the screen for 500 ms, followed by a 500 ms blank screen (i.e., inter-stimulus interval). After this, a pair of objects (stimuli A and B) remained on the screen until the participant made a response using the 'S' key for the left stimulus or the 'L' key for the correct stimulus. Stimuli A and B were presented to the left and the right of the centre of the screen, and the position of stimuli was counterbalanced across trials. Once a response had been recorded, a 250 ms inter-trial interval occurred. Participants were instructed to respond as quickly and as accurately as possible. Each participant performed four blocks in the discrimination task with no breaks between blocks. In the categorisation task, stimuli were presented in blocks, one for each

object pair continuum. At the beginning of each block, the two extremes of each object pair continuum were presented for 750 ms. Each trial began with a 250 ms fixation cross in the centre of the screen, followed by the target stimulus which remained on the screen until the participant made a response, followed by a 250 ms inter-trial interval. Participants were asked to categorise each target object as belonging to one of the two extremes presented at the beginning of each block. Each extreme was associated with the 'S' or 'L' keys. Participants were instructed to respond as quickly and as accurately as possible. Each participant performed four blocks in the Categorisation task with no breaks between blocks.

Analysis

A cut-off performance of 80% accuracy was applied in the learning phase and 9 participants did not reach the required threshold, thus their data were excluded from further analyses. Initially, we determined participants performance on the test part of the Learning phase, adopting a Kruskal-Wallis test instead of one-way ANOVA on the accuracy performance, due to violation of normality and homogeneity of variance.

First, we analysed categorisation data to estimate the category boundary for each object-pair continua, defined as the point along the morph continuum at which the stimulus value at which $P(A) = .50$, i.e., 50% probability of a Category A response. To determine whether categorical perception (CP) occurred, we compared discrimination accuracy and reaction times (RTs) for stimuli near the boundary versus stimuli at the continuum extremes. Since the extremes for each object pairing (i.e., stimuli at positions 1 and 11) had already been presented during the Learning phase, we defined images at positions 2 and 10 along each object continuum as the 'extreme' end points of each object continuum. To assess whether boundary positions differed across learning conditions, we adopted the boundary derived from the unisensory visual condition as a baseline condition, and compared it with boundaries observed in the congruent, incongruent, and irrelevant audiovisual conditions. We fitted two separated logistic mixed-effects regression (lmer)

models to the discrimination accuracy and reaction times (RTs) for stimuli at boundary vs. extremes, with fixed effects of category position and object pair. The entire dataset was subdivided into within- and between-category objects, and analyses were performed on the two different subsets separately.

To assess our second research question, we also computed the slope of the categorisation function for each object pair as an index of category sharpness. Based on each participant's individual category boundary, stimuli were classified as belonging to Category A (below the boundary) or Category B (above the boundary). The slope was defined as the difference in mean categorisation accuracy between these two sides of the continuum. A linear mixed-effects regression model was fit to slope values, with fixed effects of learning condition, object pair, and their interaction, and random intercepts for participants.

Finally, to investigate whether learning condition affected overall categorisation and discrimination performance, we separately analysed discrimination accuracy for all the object stimuli, discrimination and categorisation reaction times across learning conditions. During the discrimination task, there were two trials per object pair, thus accuracy values were 0 (stimulus never correctly identified), 0.5 (stimulus correctly identified once), and 1 (stimulus correctly identified twice). We fitted a generalized logistic mixed-effects regression (glmer) model to discrimination accuracy data with fixed effects of learning conditions and object pair. Reaction times, for both tasks, were analysed by fitting linear mixed-effects models.

The data were analysed with R via RStudio (R Core Team, 2021). Models were fitted using the 'lme4' package (Bates et al., 2014). All models adopted using the 'logit' link and the 'bobyqa' optimizer and controlled for age and sex. Post-hoc comparisons were conducted using the 'emmeans' package (Lenth Russell et al., 2022), and the Bonferroni correction was applied to correct for multiple comparisons. Experiment 1A project overview and analysis script are available

on the Open Science Framework at

https://osf.io/cwqpu/?view_only=c7d481c409c744e59bdc23664c6e0279.

2.6.2 Result

The final sample consisted of 99 participants (mean age = 33.1 years, SD = 11.4; 52.5% female), all reported to have completed secondary level education, and completed higher education (82.2%).

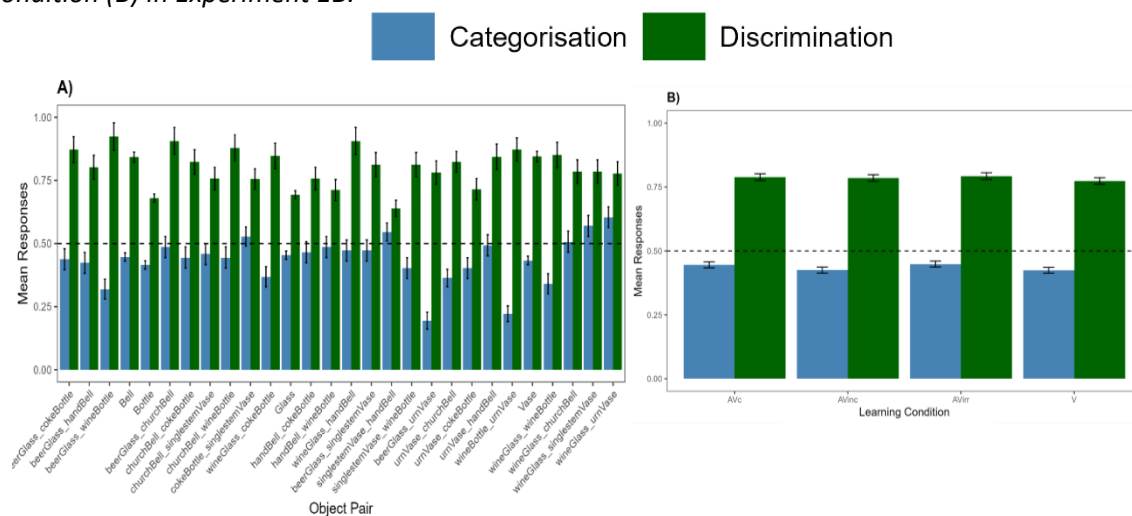
Accuracy data

Learning phase recognition data. During the Learning phase, overall participants' mean recognition accuracy for the target objects was $M = 0.95$ ($SD = 0.22$). Specifically, across each learning condition, participant mean accuracy was $M = 0.77$ ($SD = 0.42$) in the AV_c condition, $M = 0.96$ ($SD = 0.20$) in the AV_{inc} condition, $M = 0.86$ ($SD = 0.35$) in the AV_{irr} condition, and $M = 0.95$ ($SD = 0.21$) in the V condition, for mean recognition accuracy for each target stimulus refer to Supplemental Materials – Table S4. We examined participants' accuracy during the test Learning phase across the different learning conditions. Assumptions of normality (Shapiro-Wilk test: $W = 0.58$, $p < .001$) and homogeneity of variance (Levene's test: $F(3, 3052) = 62.22$, $p < .001$) were violated, therefore we adopted a non-parametric Kruskal-Wallis's test. We found no significant difference in accuracy across the learning conditions, $\chi^2(2) = 0.18$, $p = .91$.

Categorisation and discrimination data. Mean categorisation judgments and discrimination accuracy varied across the object pairs, see Figure 8A and refer to Supplemental Materials - Table S5 for mean accuracy for each target stimulus. Across all objects pairs the mean categorisation response was approximately 0.43 (SD $\approx$ 0.49). The lowest mean categorisation response was for the beerGlass - urnVase pair (M = 0.19, SD = 0.40), whereas the highest was for wineGlass - urnVase (M = 0.60, SD = 0.49), wineGlass - singlestemVase (M = 0.57, SD = 0.50), and cokeBottle - singlestemVase (M = 0.53, SD = 0.50) pairings. The overall mean discrimination accuracy across all pairs was approximately 0.81 (SD $\approx$ 0.37). The worse discrimination accuracy was observed for singlestemVase - handBell (M = 0.64, SD = 0.48), while the most accurate was found for beerGlass - wineBottle (M = 0.92, SD = 0.27) and beerGlass - churchBell (M = 0.91, SD = 0.29) pairings. Collapsing across object pairs, mean categorisation responses for each learning condition were M = 0.45 (SD = 0.50) in AV_c, M = 0.43 (SD = 0.50) in AV_{inc}, M = 0.42 (SD = 0.49) in V, and M = 0.50 (SD = 0.50) in AV_{irr}; while the mean discrimination accuracy was M = 0.79 (SD = 0.41) in AV_c, M = 0.79 (SD = 0.41) in AV_{inc}, M = 0.77 (SD = 0.42) in V, and M = 0.79 (SD = 0.41) in AV_{irr}, see Figure 8B.

Figure 8

Mean categorisation and discrimination responses by object pairing (A), and by learning condition (B) in Experiment 1B.

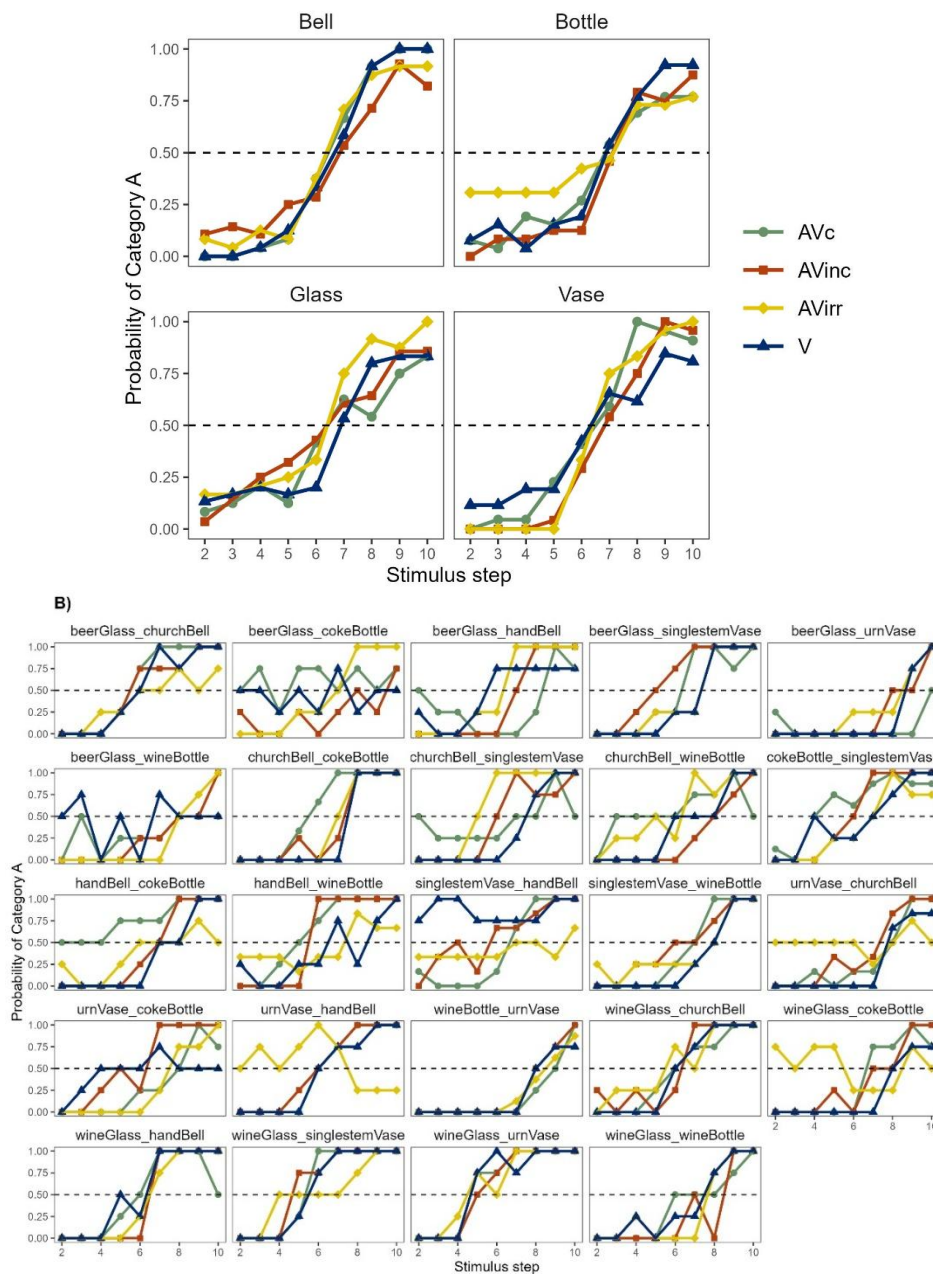


Note: Categorisation means represent the proportion of Category A responses. The plot shows categorisation (in blue) and discrimination (in green) responses by each object pairing (panel A), and by learning condition (panel B). Error bars show standard error, the dashed horizontal line shows chance level performance (0.5).

Categorisation data. The probability of categorising 'A' across the different learning conditions for each object pair is reported in Figure 9.

Figure 9

Probability of a category 'A' response across learning conditions for each object pair in Experiment 1B.

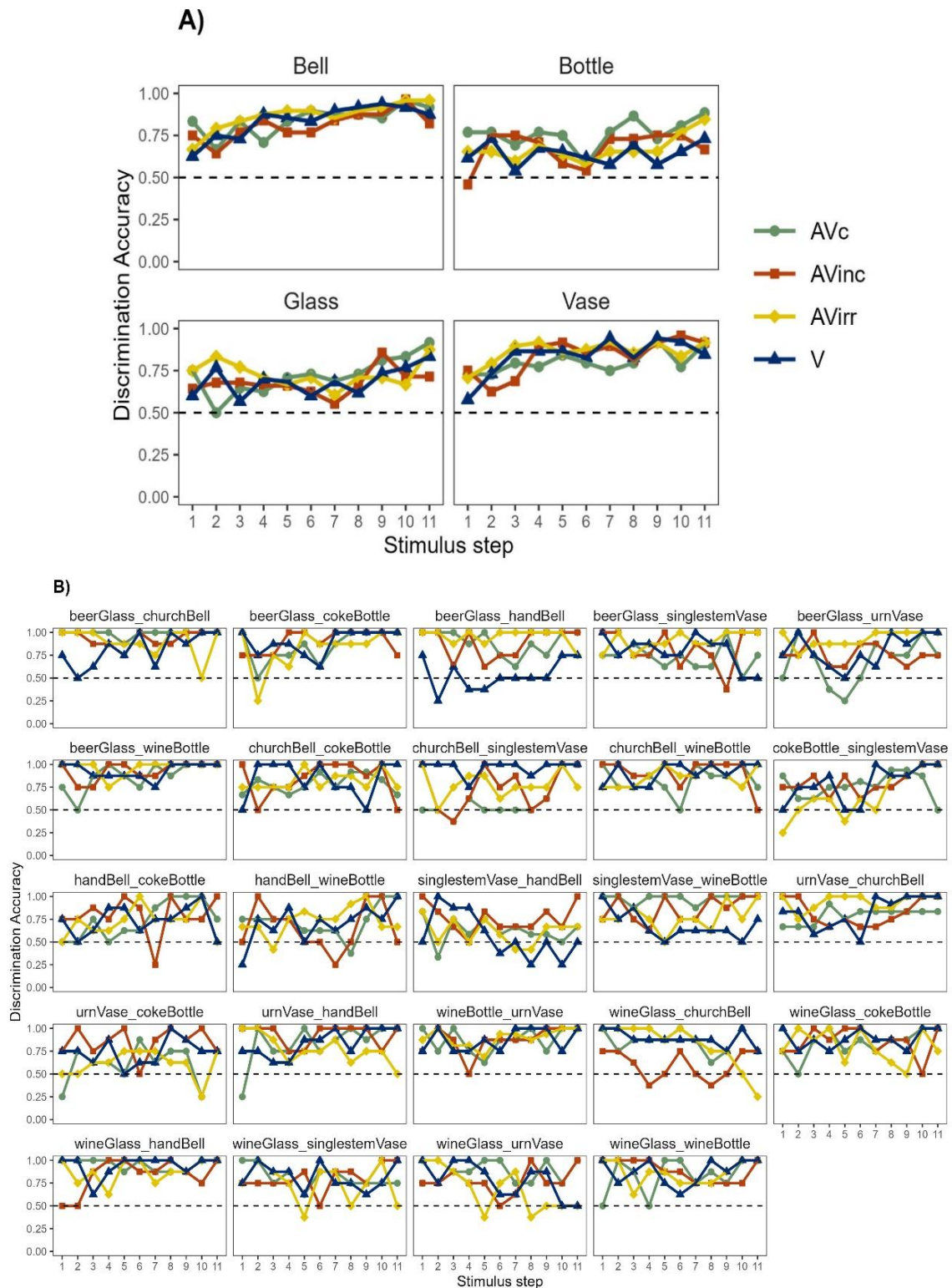


Note: The plot shows the probability of Category 'A' responses across learning conditions for within-category (panel A) and between-category object pair (panel B). The dashed horizontal line shows chance level (50%).

Discrimination data. For the within-category object pair, we compared between models, with and without the learning condition by object pair interaction. A likelihood ratio test showed a significant main effect of object pair, $\chi^2(3) = 203.2$, $p < .001$, but not of learning condition, $\chi^2(3) = 2.42$, $p = 0.49$, and it suggested that the interaction did not significantly contribute to the model predicting discrimination accuracy, $\chi^2(9) = 9.69$, $p = .38$. Post hoc comparisons showed that the odds of accurate discrimination were significantly higher for Bell compared to Bottle (*odds ratio*, $OR = 2.46$, 95% CI [1.91, 3.17], $p < .001$) and compared to Glass ($OR = 2.60$, 95% CI [2.02, 3.35], $p < .001$). In contrast, Bottle had significantly lower odds to be correctly discriminated than Vase ($OR = 0.38$, 95% CI [0.29, 0.49], $p < .001$), and as well Glass also had lower odds compared to Vase ($OR = 0.36$, 95% CI [0.28, 0.46], $p < .001$). No significant difference was observed between Bell and Vase, and between Bottle and Glass ($p = 1$). For the between-category object pair, we compared between models, with and without the learning condition by object pair interaction. A likelihood ratio test showed a significant main effect of object pair, $\chi^2(23) = 131.9$, $p < .001$, but not of learning condition, $\chi^2(3) = 3.95$, $p = .27$, and it suggested that the interaction did not significantly contribute to the model predicting discrimination accuracy, $\chi^2(9) = 5.34$, $p = .48$. Post hoc comparisons showed that the odds of accurate discrimination were significantly higher for singlestemVase - handBell ($OR = 6.06$, 95% CI [1.74, 21.09], $p < .001$) and handBell - wineBottle ($OR = 4.43$, 95% CI [1.16, 16.95], $p = .009$), compared to beerGlass – churchBell. No significant difference was observed for all the other comparisons ($p = 1$). For detailed illustration of model prediction, refer to Supplemental Materials – Figure S2.

Figure 10

Discrimination accuracy across learning conditions for each object pair in Experiment 1B.



Note: The plot shows discrimination accuracy across learning conditions for within-category (panel A) and between-category object pair (panel B). The dashed horizontal line shows chance level (50%).

Categorical perception analysis

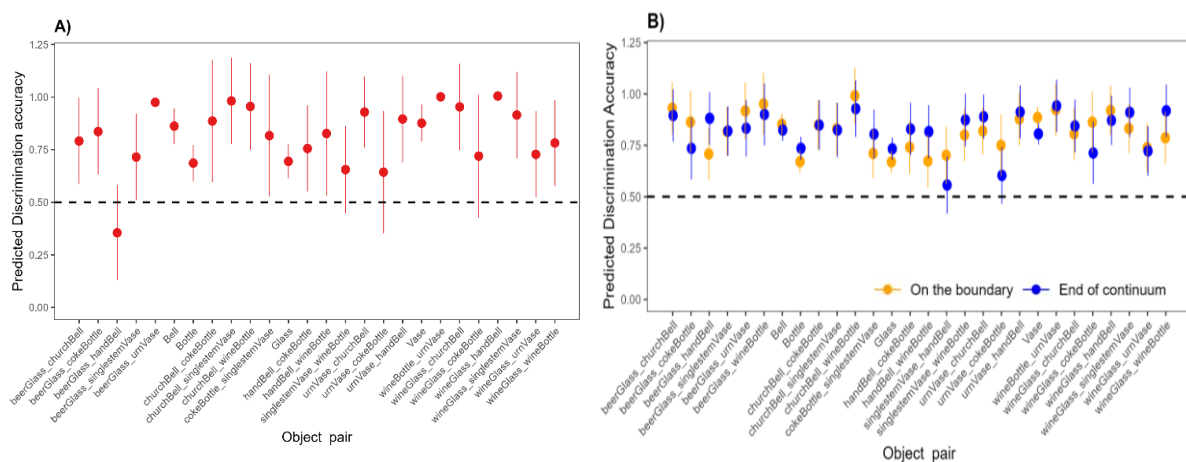
Visual-only learning condition. We first analysed the categorical and discrimination performance to the object pairs presented in the Visual-only learning condition as a baseline (and as a comparison to previous studies). For two out of 28 objects-morphed continua, participants did not show a clear category boundary across all objects, thus, they were excluded from the analysis, resulting in a total of 26 object morphed continua analysed. We identified the individual category boundary as the stimulus value at which $P(A) = .50$, i.e., 50% probability of a Category A response. Categorical perception predicts that discrimination accuracy will be higher for pairs straddling the category boundary and lower for pairs within the same category. In the data, this means that accuracy should be best at the boundary positions, especially for between-category object pairs, and be lower at the extremes for within-category object pairs. As described above, we conducted a comparison between models, with and without the object pair by position interaction. A likelihood ratio test indicated statistically significant main effects of object pair, $\chi^2(25) = 58.936, p < .001$, on discrimination accuracy, however position, $\chi^2(1) = 0.87, p < .35$, and of the object pair by position interaction did not significantly contributed to the model, $\chi^2(25) = 30.72, p = .20$, see Figure 11A. The main effects of object pair suggests that discrimination accuracy at different positions (e.g., on the boundary or at the edges) is significantly different across object pairs, which is due to the variability in the large number of object-morphed continua. Post hoc comparisons revealed that discrimination accuracy was significantly higher in the beerGlass – handBell compared to the Bell (estimated marginal means, EMM = -0.51 , 95 % CI [$-1.01, -0.01$], $p = .04$), and Vase (EMM = -0.52 , 95 % CI [$-1.02, -0.02$], $p = .03$). All other pairwise comparisons were not significant after adjustment (all $p \geq .40$).

Audio-visual learning condition. We first analyse both categorical and discrimination performance to the object pairs presented across the audio-visual learnings conditions and identified the individual category boundary. We conducted a comparison between models, with

and without the object pair by position interaction. A likelihood ratio test indicated statistically significant main effects of object pair, $\chi^2(25) = 38.1$, $p = 0.03$, and marginally significant effect of the object pair by position interaction, $\chi^2(27) = 39.6$, $p = .05$, on discrimination accuracy, however the main effects of position, $\chi^2(4) = 1.75$, $p = .78$, and learning condition, $\chi^2(3) = 0.49$, $p = .92$, did not contributed to the model, see Figure 11B.

Figure 11

The plot shows the estimated mean discrimination accuracy predicted from the linear mixed-effects model for object pairings in Experiment 1B.



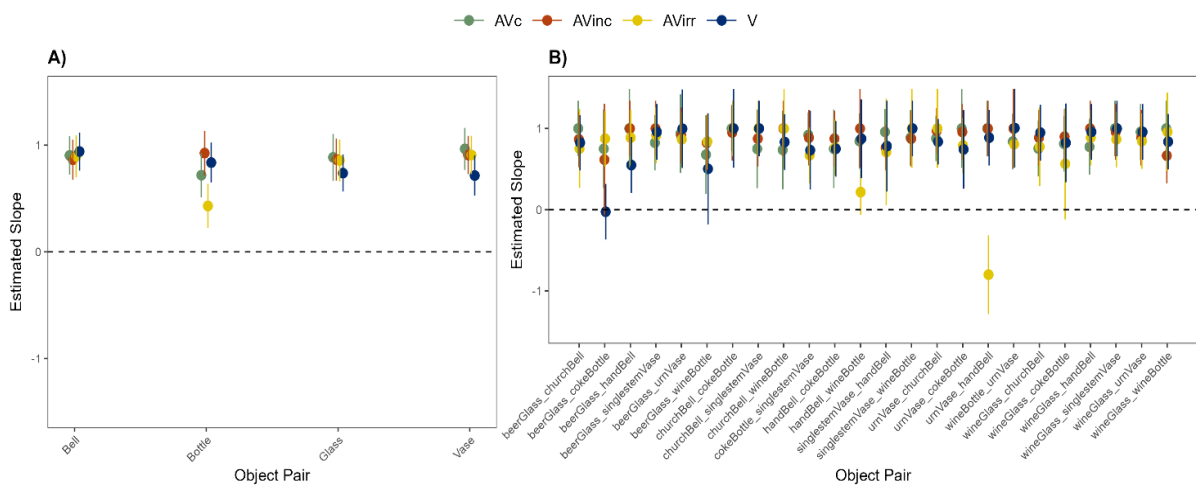
Note: Across object pairs, the plot shows discrimination estimates for the visual-only learning condition (panel A), and the audio-visual learning conditions (panel B). Error bars represent 95% confidence intervals; the dashed line indicates chance level (50%).

Slope analysis. For within category object pairs, we compared between models with and without the object pair by learning condition interaction. A likelihood ratio test reported significant main effect of learning condition, $\chi^2(12) = 23$, $p = .03$, and of learning condition by object pair interaction, $\chi^2(9) = 16.65$, $p = .05$, we found no main effect of object pair, $\chi^2(12) = 19.6$, $p = .08$. Post hoc comparisons showed that for the Bottle object pair, slopes were significantly steeper in the AV_{inc} compared to the AV_{irr} condition (EMM = 0.50, 95% CI [0.20, 0.79], $p = .03$), see Figure 12A. No other pairwise differences were significant. A second mixed-effects model tested the slopes of between-category object pairings. A likelihood ratio test reported significant learning condition by object pair interaction, $\chi^2(64) = 127.3$, $p < .001$, but no main effects of learning

condition, $\chi^2(3) = 5.77$, $p = .12$, and object pair, $\chi^2(23) = 26$, $p = .30$, see Figure 12B. Post-hoc comparisons suggested that, at the object level, for the urnVase – handBell pair, the slope was significantly steeper in both the AV_c and AV_{inc} learning conditions compared to the AV_{irr} condition (EMM = 1.8, 95% CI [0.68, 2.92], $p < .001$), and shallower slope in the AV_{irr} compared to the V condition (EMM = -1.69, 95% CI [-2.80, -0.57], $p < .001$).

Figure 12

The plot shows the estimated slope values predicted from the linear mixed-effects model for object pairings across learning conditions in Experiment 1B.



Note: Across object pairs, the plot shows slope estimates for the visual-only learning condition (panel A), and the audio-visual learning conditions (panel B). Error bars represent 95% confidence intervals, and the dashed line indicates slope equal to 0.

Reaction times data

Participant data was screened for time taken to respond to each trial (with outliers defined as ± 3 SD away from the group mean or ± 4 SD away from the individual mean, consistent with previous reports e.g. Williams et al., 2022). Within these response times, outliers were excluded from the data prior to analysis and represented of 1.2% of the entire RT data set.

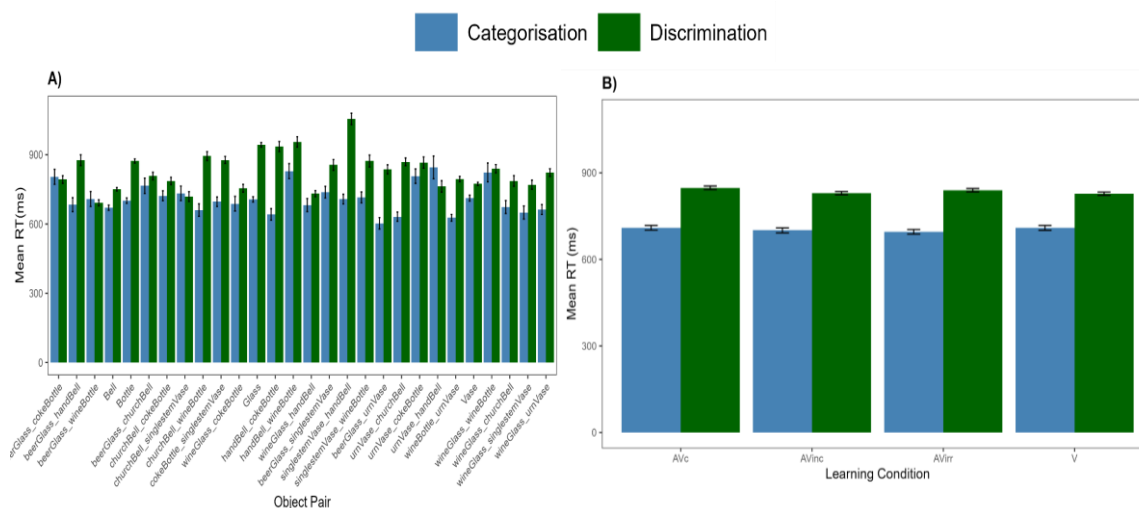
Learning phase recognition data. During the Learning phase, overall participants' mean recognition response time for the target objects was $M = 684$ ms ($SD = 439$ ms). Specifically, across

each learning condition participant's mean recognition reaction times were $M = 693$ ms ($SD = 350$ ms) in the AV_c condition, $M = 679$ ms ($SD = 257$ ms) in the AV_{inc} condition, $M = 813$ ms ($SD = 481$ ms) in the AV_{irr} condition, and $M = 685$ ms ($SD = 547$ ms) in the V condition, for mean recognition accuracy for each target stimulus refer to Supplemental Materials – Table S4. We examined participants' reaction times during the test Learning phase across the different learning conditions. Assumptions of normality (Shapiro-Wilk test: $W = 0.41$, $p < .001$) and homogeneity of variance (Levene's test: $F(3, 3052) = 4.87$, $p = .002$) were violated, thus we adopted a non-parametric Kruskal-Wallis's test. We found no significant difference in reaction times across conditions the learning conditions, $\chi^2(2) = 1.32$, $p = .52$.

Categorisation and discrimination data. Mean categorisation and discrimination reaction times varied across the object pairs, see Figure 13A and refer to Supplemental Materials - Table S6 for mean reaction times for each target stimulus. Across all object pairs, the overall mean categorisation RT was approximately 712 ms ($SD \approx 370$ ms). The fastest mean RT was observed for beerGlass – urnVase ($M = 603$ ms, $SD = 291$ ms), while the slowest mean RTs were for wineGlass – wineBottle ($M = 823$ ms, $SD = 488$ ms), handBell – wineBottle ($M = 829$ ms, $SD = 382$ ms), and urnVase – handBell ($M = 845$ ms, $SD = 575$ ms). For discrimination response times, the fastest

Figure 13

Mean categorisation and discrimination reaction times by object pairing (A), and by learning condition (B) in Experiment 1B.



mean RT was for beerGlass – wineBottle ($M = 692$ ms, $SD = 219$ ms), while the slowest was for singlestemVase – handBell ($M = 1055$ ms, $SD = 491$ ms). Overall, mean discrimination RT across all object pairs was approximately 829 ms ($SD \approx 340$ ms). Mean reaction times in the categorisation task for each learning condition were $M = 709$ ms ($SD = 347$ ms) in AV_c , $M = 701$ ms ($SD = 364$ ms) in AV_{inc} , $M = 709$ ms ($SD = 335$ ms) in V , and $M = 696$ ms ($SD = 337$ ms) in AV_{irr} condition. The mean reaction times in the discrimination task was $M = 848$ ms ($SD = 364$ ms) in AV_c , $M = 830$ ms ($SD = 347$ ms) in AV_{inc} , $M = 827$ ms ($SD = 353$ ms) in V , and $M = 840$ ms ($SD = 357$ ms) in AV_{irr} condition, see Figure 13B.

Note: The plot shows categorisation (in light blue) and discrimination (in green) reaction times in milliseconds by each object pairing (panel A), and by learning condition in (panel B). Error bars show ± 1 standard error across participants.

Categorisation reaction times. We compared between models, with and without the learning condition * object pair interaction. A likelihood ratio test suggested significant main effect of object pair, $\chi^2(27)=52.7$, $p=0.002$, and that the interaction did significantly contribute to the model predicting categorisation reaction times, $\chi^2(81) = 122.6$, $p < .001$, while the main effect of learning condition was not significant, $\chi^2(3) = 1.73$, $p=.63$. Post-hoc comparisons suggested that for the beerGlass – cokeBottle pair, participants categorise the stimuli within the continuum significantly faster when previously learned in the AV_c (EMM = -1.61 ms, 95% CI [-2590, -627], $p < .001$), in the AV_{inc} (EMM = -1.39 ms, 95% CI [-2369.9, -413], $p < .001$), and the AV_{irr} (EMM = -1.78 ms, 95% CI [-2759.4, -803], $p < .001$) compared to when objects had been previously learned in the V condition.

Discrimination reaction times. We compared between models with and without the learning condition by object pair interaction. A likelihood ratio test suggested that the learning condition $\times$ object pair interaction did significantly contribute to the model predicting discrimination accuracy ($\chi^2(69) = 228.61$, $p < .001$). Post hoc comparisons suggested that,

irrespective of learning modality, participants were significantly faster at discriminating Bell compared to both Bottle (EMM = -171 ms, 95% CI [-253, -90], $p < .001$), and Glass (EMM = -196 ms, 95% CI [-276, -116], $p < .001$), and slower at discriminating between Bell (EMM = 227ms, 95% CI [145, 309], $p < .001$) and Glass (EMM = 251 ms, 95% CI [170, 333], $p < .001$) compared to Vase. Furthermore, for the beer Glass – hand Bell pair, participants were significantly faster to discriminate the stimuli when previously learned in the AV_c (EMM = -1323 ms, 95% CI [-2156, -500], $p < .001$), in the AV_{inc} (EMM = -1215 ms, 95% CI [-2064, -366], $p < .001$), and the AV_{irr} condition (EMM = -1042 ms, 95% CI [-1883, -202], $p < .001$) compared to when stimuli had been previously learned in the V condition. All other comparisons were not significant ($p = 1$).

Categorical perception analysis

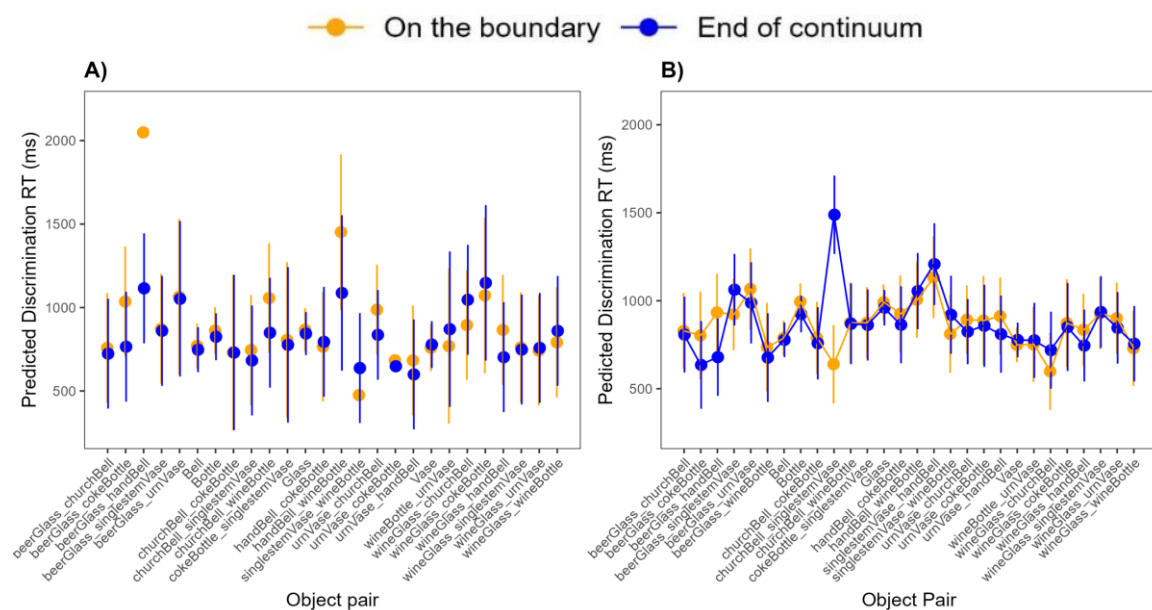
Visual-only learning condition: We compare between models, with and without the object pair by position interaction. A likelihood ratio test found significant main effects of position, $\chi^2(26)=59.71$, $p<.001$, of object pair, $\chi^2(50)=103.6$, $p<.001$, and their interaction, $\chi^2(25)=55.95$, $p<.001$, see Figure 14A. Post hoc comparisons for the interaction showed that, for the beerGlass–handBell pair, participants were significant slower at discriminating stimuli straddling the category boundary compared to stimuli at the ends of the continuum (estimated difference = 933.8 ms, CI 95% [473, 1394], $p<.001$), with all other comparisons reporting not significant p values (all $p \geq .99$).

Audio-visual learning condition: A comparison between models, with and without the learning condition by position interaction, was performed, and a likelihood ratio test suggested that interaction did not significantly contribute to the model predicting discrimination reaction times, $\chi^2(3) = 4.13$, $p = .25$. We found significant main effects of both position, $\chi^2(28) = 54.6$, $p = .002$, object pair, $\chi^2(25) = 71.4$, $p < .001$, and their interaction, $\chi^2(27) = 54.6$, $p = .001$, while we found no significant main effect of learning condition, $\chi^2(3) = 2.76$, $p = .4$, see Figure 14B. Post hoc comparisons for the interaction showed that, for the churchBell - singlestemVase, participants

were significant faster at discriminating stimuli straddling the category boundary compared to stimuli at the ends of the continuum (estimated difference = -849.53 ms, 95% CI $[-1274, -425]$, $p < .001$), with all other comparisons reporting not significant p values ($p = 1$).

Figure 14

Plots showing estimated mean discrimination reaction times predicted from the linear mixed-effects model for object pairings and position In Experiment 1B.



Note: Across object pairs, the plot shows discrimination estimates for the visual-only learning condition (panel A) and the audio-visual learning conditions (panel B). Error bars represent 95% confidence intervals.

2.6.3 Discussion

The results of Experiment 1B suggest that objects are not categorically perceived, both when have been learned visually-only or paired with sounds. Moreover, semantically congruent sounds during learning did not benefit overall subsequent categorical perception: specifically, we did not observe a significant sharpening of the category boundary in the categorisation task, or an increased discriminability for stimuli straddling the cross-category boundary compared to the extremes of each object continuum. Moreover, we fail to find evidence for CP even in the visual-only learning condition (contrary to previous reports that have demonstrated categorical

perception effects for visually learned object continua, e.g., Goldstone, 1994b; Livingston et al., 1998; Newell & Bülthoff, 2002; Op de Beeck et al., 2003). These findings might be due to a variety of factors. First, our results might be related to the inability of sounds (or the multisensory pairing of sounds and images) to capture individual variability in different object representations. Second, the paradigm structure itself might elicit different cognitive processes: while the learning phase may be based on an old-new recognition task related to memory and attention, the categorical perception phase is more related to perceptual processes. As such, training one cognitive process might not necessarily benefit another separate cognitive process. Murray and colleagues found that a single exposure of multisensory object stimuli can benefit subsequent unimodal identification of the same objects (Lehmann & Murray, 2005; Matusz et al., 2019; Thelen & Murray, 2013; Murray & Sperdin, 2010; Tovar et al., 2020). However, they adopted a continuous task paradigm in which the time distance between ‘exposure’ and ‘discrimination’ tasks was a few hundred milliseconds. Instead, in Experiment 1B we adopted a paradigm in which the time between ‘learning’ and ‘categorical perception’ phases was approximately of five minutes. This larger time interval might have significantly reduced any possible enhancement in participant performance. Furthermore, according to the late integration hypothesis (Rohe & Noppeney, 2015), sounds may enhance discrimination performance when presented during the decisional phase of the task, rather than during the perceptual phase. Finally, the Learning phase in Experiment 1B is a simple exposure task, in which participants had to passively perceive both multisensory and unisensory stimuli, without any type of supervised learning (i.e., feedback). This, combined with possibly an insufficient number of trials in the Learning phase, might have resulted in a reduction in any possible benefit to the object representations.

Overall, participants were able to categorise objects into distinct categories, especially for within- class of objects. However, we did not observe improved discriminability at the boundary. Several factors may account for this null effect. One possibility is that limitations in the morphing procedure used to generate the object continua: despite efforts to ensure perceptually uniform

transitions, non-linear or uneven changes in shape across the morph space may have reduced sensitivity to the object categorical structure. Another aspect might be the nature of the auditory stimuli, that while semantically congruent at the object level, may have lacked sufficient specificity or perceptual salience to strongly reinforce object identity during learning. Results from Experiment 1A shows high variability in the sound–object matching performance, especially for Glass and Vase categories, suggesting that the semantic relation and perceptual distinctiveness of the auditory stimuli may have weakened the strength of multisensory learning. Thus, the absence of enhanced categorical perception following congruent audiovisual learning in Experiment 1B, might at least in part reflect insufficiently robust auditory representations for certain object categories. For example, if Glass sounds were inconsistently perceived due to individual variability in category prototypes or a mismatch between static images and dynamic sound cues, then the effectiveness of multisensory learning in that condition may have been reduced. Similarly, the poor match between Vase sounds and both category and exemplar representations suggest that those auditory cues may have failed to meaningfully support visual category learning. These outcomes indicate that the task-relevance and semantic specificity of auditory cues are critical for multisensory learning effects to emerge (Bizley et al., 2016; Roark et al., 2019).

Moreover, we also found a difference in the overall reaction time for categorisation and discrimination tasks: irrespective of learning condition, participants were faster at discriminating between Bell and Vase objects. This may reflect the increased cognitive demands of the categorisation task, which required participants to apply learned category boundaries to novel stimuli. In contrast, the discrimination task relied more on perceptual comparison between stimuli presented on the same trial, likely allowing for faster, more stimulus-driven responses. An overall RT benefit of the audio-visual learning condition, irrespective of whether sounds were congruent, incongruent, or irrelevant, was also found for beer Glass – hand Bell during the discrimination task and for beer Glass – coke Bottle for discrimination performance.

2.7 Experiment 1C

In Experiment 1C, we investigated the perceptual similarity of the morphed object continua adopted in Experiment 1B, to verify that perceptual similarity judgments increased monotonically with morphing level, and that intermediate stimuli were judged more similar to both category endpoints than to unrelated categories. Thus, we wanted to validate that the object continua supported reliable estimates of category boundaries and slopes in Experiment 1B. Perceived similarity has been proposed as a central mechanism underlying category learning (Goldstone, 1994b; Ashby & Maddox, 2005): categorisation performance is influenced by the psychological distance between exemplars, such as the more similar a stimulus is to a known category member, the more likely it is to be categorised accordingly. Based on similarity-based categorisation models (Nosofsky, 1986; Shepard, 1987), which propose that classification decisions depend on perceived psychological similarity between stimuli, we expected that object pairs sampled from the same morph continuum would be judged as increasingly dissimilar with greater distance from one another. By collecting similarity ratings for pairs of morphed objects, Experiment 1C allowed us to assess whether perceived similarity corresponded systematically to morph position, particularly between the endpoints and intermediate stimuli. Participants performed a visual similarity rating task in which they judged the perceived similarity between object pairs sampled across the object continua.

2.7.1 Method

Participants

Based on a priori power analysis performed on PANGEA (Westfall, 2016; v0.2), a sample size of 45 participants was calculated as required in order to identify a two-way within-factors repeated measures ANOVA effect, with *step distance* (4 levels) and *object category* (2 levels) as factors, with an effect size (f) of 0.25, $\alpha = 0.05$, and power ($1 - \beta$) = 0.80. We recruited fifty-three adult participants (mean age = 41.47 years, SD = 12.66, 50% female) using Prolific

(<https://www.prolific.co/>), based on the following inclusion criteria: fluency in English, normal or corrected-to-normal vision and no hearing impairments. Data from thirteen participants was excluded (five did not complete the experiment, and seven did not fit the inclusion criteria). The final analysis sample consisted of 41 participants. All participants were naïve to the purpose of the study and were paid for participating (at a rate of £9.00 per hour). The study was approved by the Trinity College Dublin, School of Psychology Research Ethics Committee and complied with EU data protection legislation (GDPR).

Stimuli

The visual stimuli used in Experiment 1C are those described in the General Stimuli. Specifically, we selected four within-category and two between-category continua. The list of objects morphed continua (including the two exemplars) included the following: Church-Hand Bell, Wine-Coke Bottle, Wine-Beer Glass, Urn-Single Stem Vase, Coke Bottle – Single Stem Vase, Urn Vase – Wine Bottle.

Design

Experiment 1C was based on a two-way, fully factorial within-subjects design, with the distance between image pairs along each morph continuum as one factor (image step distance; 4 levels) and object category (between and within, 2 levels) as factors. Participants performed a similarity rating task on displayed pairs of images using a 7-point Likert scale. Although each image was compared to another differing by 1–4 steps, the focus of the analysis was on the stepwise difference rather than the individual images themselves. After a training phase, participants were instructed to judge how similar each image of object continuum, presented side by side were to each other. The presentation order of stimuli was randomised in each trial (e.g., stimulus A on the left or right, and the same for stimulus B). We manipulated the distance between stimulus A and B for each object pair, i.e., 4 steps, 3 steps, 2 steps, or 1 step apart to assess at which point along the morphed continua the objects were perceived as more similar compared to the others. The

dependent variables in this experiment were similarity ratings and reaction times. Each continuum was blocked, and block order was counterbalanced across participants, and the presentation order of trials within each block was fully randomised. This study had a total of 476 ($34 \times 2 \times 7$, n of comparison $\times n$ of repetitions $\times n$ of object continuum) experimental trials. Participants took approximately 25-30 minutes to complete the study.

Procedure

Participants provided informed consent and completed basic demographic questionnaire (see Participant section). After a short training phase (using stimuli not presented in the main task), the main task began. On each trial, a fixation cross appeared at the centre of the screen for 250 ms, followed by a pair of object images presented simultaneously to the left and right of the centre of the screen. The stimuli remained on the screen until a response was made using the mouse, followed by a 500 ms inter-trial interval. The position of each image (i.e., left or right) was randomised on every trial to control for side bias. Participants were instructed to examine the pair of object images and rate how similar they were to each other on a 7-point Likert scale, ranging from 'very dissimilar' (1) to 'very similar' (7). Participants were informed to use the whole scale to express their judgements. Feedback was provided, '*Too slow! Respond faster in the next trial*', if responses times were longer than three seconds.

Data Analysis

First, we normalised participants' ratings by transforming each individual data point (e.g., similarity ratings) to a range between 0 and 1 using the Min-Max Normalisation function (Patro & Sahu, 2015). To assess whether perceived similarity between images decreases as a function of morph distance, and if this pattern differs for within-class versus between-class object pairs, we conducted linear mixed-effects regression (lmer) models to the normalised similarity ratings, with steps, object category, and their interaction as fixed effects, controlling for age and sex. We included participant and image pair as random intercepts, to account for potential variability due

to specific images or image pairs. Although our primary analyses focused on similarity ratings, we also examined reaction times as an exploratory measure to assess potential differences in response speed across steps. Moreover, for each pair of stimuli, we computed the mean similarity rating across participants separately for within- and between- category trials, and plotted this data as heatmaps, separately for within- and between-category. Finally, we calculated difference scores across within- and between-category similarity at the item-pair level. Specifically, for each A–B pair we subtracted the mean between-class similarity from the mean within-class similarity at the corresponding positions. This visualisation allows to explore how the influence of morph distance differs between within- and between-category pairs.

The data were analysed with R via RStudio (R Core Team, 2021). Models were fitted using the ‘lme4’ package (Bates et al., 2014). Post-hoc comparisons were conducted using the ‘emmeans’ package (Lenth Russell et al., 2022), and the Bonferroni correction was applied to correct for multiple comparisons. Experiment 1C project overview and analysis script are available on the Open Science Framework at https://osf.io/b3yrm/?view_only=ff927c6e28324e9a8d90b999d9eae52.

2.7.2 Result

Similarity ratings data

We assessed whether participants responded using the whole rating scale, participants that used less than one third of the scale (i.e. only choosing one answer consistently) were excluded from further analysis. Seven participants were excluded due to lack of effort (e.g., not using the whole rating scale), resulting in a final sample of 41 participants (mean age = 26.6 years, SD = 9.85; 39% female). The majority (68.3%) reported having completed formal education, and 78.1% held a third-level qualification or higher. Mean similarity ratings decreased across the four comparison steps: the mean of normalised similarity rating was 0.76 (SD = 0.25) for one step distance, 0.6 (SD = 0.27) for two steps distance, 0.49 (SD = 0.28) for three steps distance, and 0.42

(SD = 0.28) at four steps distance. For means similarity ratings across steps for each object category see Table 1.

Table 3

Mean similarity ratings and reaction times (in milliseconds) across steps for each object category in Experiment 1C.

Object pair	Step	Reaction times	Normalised Similarity Ratings	
		Mean (ms)	Mean	SD
Bell	1	840	0.51	0.30
	2	756	0.56	0.26
	3	672	0.45	0.26
	4	588	0.37	0.26
Bottle	1	860	0.80	0.23
	2	774	0.72	0.24
	3	688	0.62	0.26
	4	602	0.57	0.26
Coke Bottle – stem Vase	1	820	0.73	0.25
	2	738	0.61	0.27
	3	656	0.50	0.26
	4	574	0.43	0.25
Glass	1	840	0.77	0.24
	2	756	0.65	0.26
	3	672	0.55	0.29
	4	588	0.51	0.30
Urn Vase – wine Bottle	1	840	0.74	0.27
	2	756	0.53	0.29
	3	672	0.41	0.28
	4	588	0.33	0.26
Vase	1	840	0.73	0.25
	2	756	0.53	0.26
	3	672	0.40	0.26
	4	588	0.32	0.26

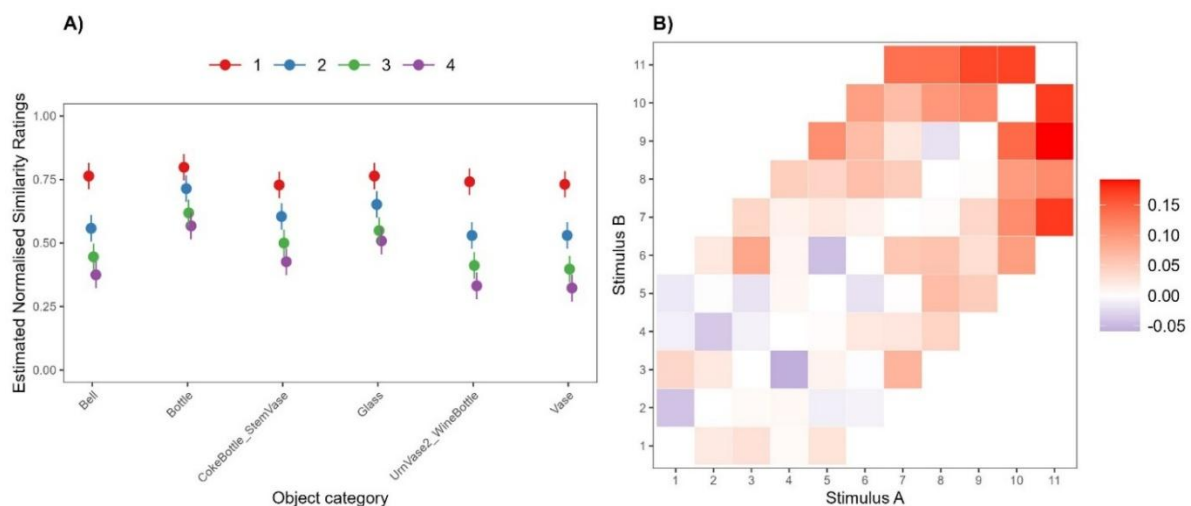
A comparison between models, with and without the object category * step interaction, was performed on the normalised similarity ratings. A likelihood ratio test suggested a significant object category * step interaction effects, $\chi^2(15) = 273.56$, $p < .001$, as well as significant main effects of both object category, $\chi^2(5) = 1142.5$, $p < .001$, and step, $\chi^2(3) = 4501.2$, $p < .001$, see Figure 15A. Post hoc contrasts suggested that similarity ratings decreased as the morphing step distance increased.

For within-category object, pairs of Bell stimuli were rated as significantly more similar at one step distance compared to pairs with two steps (estimated marginal means, EMM = 0.21, 95% CI [0.17, 0.25], $p < .001$), at three steps (EMM = 0.33, 95% CI [0.29, 0.37], $p < .001$), at four steps (EMM = 0.40, 95% CI [0.36, 0.44], $p < .001$) distance. Pairs of stimuli at two step distance were rated as significantly more similar compared to pairs with three steps (EMM = 0.12, 95% CI [0.08, 0.16], $p < .001$) and four steps (EMM = 0.19, 95% CI [0.15, 0.23], $p < .001$) distance. Finally, pairs of stimuli at three steps distance were rated as significantly more similar compared to pairs with four steps distance (EMM = 0.07, 95% CI [0.03, 0.12], $p < .001$). For Bottle, one-step pairs were rated as significantly more similar than two steps (EMM = 0.08, 95% CI [0.05, 0.12], $p < .001$), three steps (EMM = 0.19, 95% CI [0.15, 0.23], $p < .001$), and four steps (EMM = 0.24, 95% CI [0.20, 0.28], $p < .001$). Two-step pairs were rated as significantly more similar than three steps (EMM = 0.10, 95% CI [0.06, 0.14], $p < .001$) and four steps (EMM = 0.16, 95% CI [0.11, 0.20], $p < .001$). Three-step pairs were rated as significantly more similar than four steps (EMM = 0.05, 95% CI [0.01, 0.10], $p = .003$). For Glass, one-step pairs were rated as significantly more similar than two steps (EMM = 0.11, 95% CI [0.08, 0.15], $p < .001$), three steps (EMM = 0.22, 95% CI [0.18, 0.26], $p < .001$), and four steps (EMM = 0.26, 95% CI [0.22, 0.31], $p < .001$). Two-step pairs were rated as significantly more similar than three steps (EMM = 0.11, 95% CI [0.07, 0.15], $p < .001$) and four steps (EMM = 0.15, 95% CI [0.11, 0.19], $p < .001$). However, the difference in ratings between three and four steps was not significant (EMM = 0.04, 95% CI [-0.00, 0.09], $p = .069$). For Vase, one-step pairs were rated as significantly more similar than two steps (EMM = 0.20, 95% CI [0.17, 0.24], $p < .001$), three steps (EMM = 0.34, 95% CI [0.30, 0.38], $p < .001$), and four steps (EMM = 0.42, 95% CI [0.38, 0.46], $p < .001$). Two-step pairs were rated as significantly more similar than three steps (EMM = 0.14, 95% CI [0.10, 0.18], $p < .001$) and four steps (EMM = 0.22, 95% CI [0.17, 0.26], $p < .001$). Three-step pairs were also rated as significantly more similar than four steps distance (EMM = 0.08, 95% CI [0.03, 0.12], $p < .001$).

For between-category object, pairs of CokeBottle–StemVase stimuli were rated as significantly more similar at one-step compared to two steps (EMM = 0.13, 95% CI [0.09, 0.17], $p < .001$), three steps (EMM = 0.24, 95% CI [0.20, 0.28], $p < .001$), and four steps (EMM = 0.31, 95% CI [0.27, 0.36], $p < .001$) distance. Two-step pairs were significantly more similar than three steps (EMM = 0.11, 95% CI [0.07, 0.15], $p < .001$) and four steps (EMM = 0.19, 95% CI [0.15, 0.23], $p < .001$). Three-step pairs were also more similar than four steps (EMM = 0.08, 95% CI [0.03, 0.12], $p < .001$). Finally, for pairs of UrnVase–WineBottle stimuli, one-step pairs were rated as significantly more similar than two steps (EMM = 0.22, 95% CI [0.18, 0.26], $p < .001$), three steps (EMM = 0.34, 95% CI [0.30, 0.38], $p < .001$), and four steps (EMM = 0.43, 95% CI [0.38, 0.47], $p < .001$). Two-step pairs were rated as significantly more similar than three steps (EMM = 0.12, 95% CI [0.08, 0.16], $p < .001$) and four steps (EMM = 0.21, 95% CI [0.17, 0.25], $p < .001$). Three-step pairs were also rated as significantly more similar than four steps (EMM = 0.08, 95% CI [0.04, 0.13], $p < .001$).

Figure 15

The plot shows normalised similarity ratings responses for Experiment 1C.



Note: The plot in panel A) shows the interaction effect of object category by step distance.

Error bars represent 95% CIs. The plot in panel B) displays a heatmap of difference scores, where the x- and y-axes represent the two objects in each pair, and the colour scale reflects the mean difference score (blue = negative, red = positive). For each A–B pair, the mean normalised ratings

between-category was subtracted from the mean ratings of within-category pairs at the corresponding positions, thus positive values indicate that within-category pairs were rated as more similar than between-category pairs, whereas negative values indicate the opposite.

Reaction Time data

Participant data was screened for time taken to respond to each trial (with outliers defined as ± 3 SD away from the group mean or ± 4 SD away from the individual mean, consistent with previous reports e.g. Williams et al., 2022). Within these response times, outliers were excluded from the data prior to analysis and represented of 16.9% of the entire RT data set. Mean reaction times were 249 ms at one step, 232 ms at two steps, 247 ms at three steps, and 221 ms at four steps. A comparison between models, with and without the object pair by step interaction, was performed and a likelihood ratio test suggested that the interaction did not significantly contribute to the model predicting the probability of faster reaction times ($\chi^2(6) = 0.3$, $p = 0.58$) across object pairs and different steps.

2.7.3 Discussion

The results of Experiment 1C showed that perceived similarity between images decreased systematically as morph distance increased along the object continuum, supporting the prediction that greater visual distance corresponds to lower similarity judgments. This effect was observed for both within- and between-category object pairs, with within-category pairs consistently rated as more similar than between-category pairs, as reflected in the difference scores (Figure 15B). The object category by step interaction was significant, indicating that the influence of morph distance on similarity ratings varied across object categories. Main effects further confirmed that similarity ratings decreased monotonically with increasing step distance and differed across objects. Post hoc analyses further revealed that a stepwise decline pattern in similarity ratings within each object category, with significant differences between adjacent and nonadjacent steps for nearly all objects, except for Glass at Step 3–Step 4 ($p = .069$). In contrast, reaction times were

consistent across steps and object categories, suggesting that response speed was not influenced by morph distance or different object category.

Overall, these findings confirm that participants perceived stimuli as progressively less similar with increasing morph distance and that this pattern was modulated by object category. Importantly, the consistent similarity ratings across steps and categories validate the morphed stimulus set adopted in Experiment 1A, demonstrating that it effectively captures perceptual and categorical structure in visual similarity judgments.

2.8 General Discussion

In Chapter 2 we investigate whether multisensory prior information could influence object categorisation. Specifically, we examine whether auditory information influences the visual categorisation of familiar objects. At this end, we developed a new paradigm focusing on the role of prior multisensory (audio-visual) exposure on subsequent visual categorical perception performance. We investigated the appropriateness of auditory stimuli related to the corresponding visual object categories, and the perceptual similarity of visual morphed objects to the continuum extremes. Despite previous findings of multisensory benefit on subsequent unimodal discrimination performance (Lehmann & Murray, 2005; Matusz et al., 2019; Thelen & Murray, 2013; Murray & Sperdin, 2010; Tovar et al., 2020), we did not observe an overall multisensory enhancement on later visual categorisation and discrimination accuracy. Additionally, contrary to previous studies on unimodal categorical perception of familiar objects founding evidence for categorical perception in a similar set of object stimuli set (Newell & Bülthoff, 2002), we did not find compelling evidence for categorical perception (CP) across objects. For the purposes of this discussion, we will address these two aspects separately: the absence of a multisensory benefit will be discussed, followed by the absence of evidence for CP.

The absence of a multisensory benefit

The results of Experiment 1B failed to show an advantage of semantically congruent sounds during learning on subsequent visual categorical perception of familiar objects. Specifically, we did not observe significant sharpening of the category boundary as we had predicted, nor did we observe an enhanced discriminability of object images straddling the perceived cross-category boundary. There are multiple reasons that may explain these results, for instance the structure of the paradigm adopted, the necessary conditions for multisensory benefit to occur, and the nature of the auditory stimuli themselves.

The paradigm structure. The paradigm adopted in Experiment 1B targeted different cognitive processes through its two phases (i.e., Learning and Categorical Perception phase). While the Learning phase assessed memory and attention through an old-new recognition task, the Categorical Perception phase investigated perceptual performance on discrimination and categorisation tasks. Therefore, the short unsupervised training task in the Learning phase might not necessarily advantage another different cognitive process. Moreover, the multisensory benefit reported by Murray et al. (2010) is interpreted through the ‘momentary identification hypothesis’ of the conceptual short-term memory model (CSTM; Intraub, 1980; Potter, 1976). The CSTM, initially developed for investigations of the visual domain only, suggests that a memory buffer quickly integrates (i.e., momentary identification) new perceptual input with the one already stored in long-term memory, building new representations stored in long-term memory. Murray and colleagues (e.g., Murray et al., 2005; Thelen et al., 2012) further expand the CSTM to integrate multisensory information, suggesting that the integration of semantically congruent audiovisual input can enhance the encoding and consolidation of object representations (Murray et al., 2005; Thelen et al., 2012; 2013). According to this view, exposure to congruent multisensory stimuli supports richer and interconnected memory traces, facilitating effective retrieval even when only one modality is later available. In this context, congruent audiovisual pairings may promote greater encoding, increasing the likelihood of subsequent unisensory recognition by engaging both perceptual and conceptual memory systems during initial learning. The extended

CSTM is suggested to utilise as well auditory information in conjunction with visual information: improved visual identification performance is reported for semantically congruent AV stimuli and, and it is worsened for semantically incongruent stimuli (Chen & Spence, 2010). Thus, audio-visual semantic interactions take place in a short-term buffer, that quickly allows access semantic representations to produce a coherent multisensory object representations, that in turn aid identification or discrimination (Chen & Spence, 2010). In tasks involving the rapid presentation of objects, visual stimuli are briefly recognised before being rapidly overwritten by subsequent stimuli. This interpretation/principle applies to the continuous task paradigm used by Murray et al. (2005), in which initial and repeated presentations of the same stimuli occur within an interleaved stream of stimuli. The paradigm is characterised by only a few milliseconds between the ‘exposure’ and ‘discrimination’ tasks, thereby probing the potential multisensory benefits on ‘trace memories’ (Potter, 1993) in short-term memory. In the paradigm used in Experiment 1B, the interval between the two phases was determined by a self-timed break lasting between three and five minutes. This extended interval may have been sufficient to dissipate any potential multisensory benefits for discrimination and categorisation performance. Moreover, as noted earlier, the Learning phase was unsupervised, allowing participants to associate specific auditory stimuli with corresponding visual object images. However, the number of trials in this phase may have been insufficient to produce robust activation of object representations. The main effect observed was faster discrimination at the object-specific level (e.g., beer Glass – hand Bell vs. beer Glass – coke Bottle). This benefit was found in all three audio-visual learning conditions relative to a single visual one, irrespective of whether sounds were congruent, incongruent, or irrelevant.

Although a multisensory benefit on subsequent unimodal identification has been found using the similar paradigm in numerous experiments across different domains—including behavioural (Laurienti et al., 2004; Molholm et al., 2006), ERP (Murray et al., 2005; Murray & Sperdin, 2010; Thelen et al., 2012), and fMRI studies (Murray et al., 2005; Gleiss & Kayser, 2014)—other research has failed to replicate this effect or has found more limited evidence depending on

task or stimulus type (Doehrmann & Naumer, 2008; Stevenson et al., 2013). In studies adopting a slightly different paradigm aimed at investigating memory processes more deeply (e.g., encoding and retrieval), there is little evidence of benefit or disadvantage due to congruent or incongruent multisensory stimuli respectively on subsequent unimodal identification (Heikkilä et al., 2015; Canits et al., 2018; Moran et al., 2013; Nyberg et al., 2000). For instance, Heikkilä et al. (2015), adopting a two-part recognition memory task, found improved recognition of object sounds when presented with semantic congruent images, but no overall enhancement of visual recognition for object presented with semantically related sounds. Additionally, they found that information from the irrelevant modality still influenced memory performance, suggesting that semantically congruent stimuli in the unattended modality were also processed at a semantic level, despite not being explicitly attended (Heikkilä et al., 2015). Canits et al. (2018) argued that the absence of evidence for multisensory benefits may reflect the fact that conditions facilitating initial stimulus processing do not necessarily translate into improved memory performance for that specific stimulus. In an adapted stimulus–response grasp compatibility task, Canits et al. (2018) found that congruent motor actions facilitated immediate semantic categorisation but did not enhance long-term memory for object labels or images in both object recognition and free recall tasks (Canits et al., 2018). The lack of evidence for multisensory memory enhancement suggests that multisensory- or motor–semantic congruency does not necessarily translate into improved long-term memory representations. Another study failed to replicate the multisensory benefit: Nyberg et al. (2000) reported no benefit in recall accuracy for visual words previously encoded in a multisensory manner compared to words encoded as unimodal stimuli, although multisensory encoding was related to increased activation of the right auditory cortex in response to visual word (Nyberg et al., 2000). Even though the tasks adopted in the above-mentioned studies are different from the learning (exposure and test) task described in this Chapter, there is no clear explanation as to why the multisensory benefit should be limited to the continuous recognition task only.

The necessary conditions for multisensory benefit to occur. Multisensory benefit might be task-specific and not due to object-based semantic congruency per se. For instance, Pecher & Zeelenberg (2022), adopting the continuous recognition task originally developed by Thelen et al., (2015), failed to find compelling evidence for multisensory (AV) benefit in object recognition, compared to unimodal (i.e., as images or sounds) prior presentation, and reported a strong memory benefit when the same multisensory stimuli were repeated at test task. These findings suggest that multisensory encoding did not provide a general memory advantage. Notably, they observed benefit in performance only for task-relevant auditory information presented during retrieval, which suggests that the multisensory benefits may depend on the matching between encoding and retrieval modalities, rather than the richness of multisensory encoding per se (Pecher & Zeelenberg, 2022). Thus, the multisensory memory benefit previously found might be task-specific, instead of universal finding: a response bias or attention induced effect associated with the nature of the task itself may have played a role in improving memory. The paradigm adopted in Experiment 1B focused on effect of prior multisensory exposure on later visual discrimination and categorisation. In this design, the sound was task-irrelevant, which may partly explain the absence of clear evidence for a multisensory benefit on subsequent categorisation, which rely on memory (Murphy, 2004). Finally, in Experiment 1B, the Learning phase may not have sufficiently activated individual object knowledge due to paradigm limitations or participant-specific variability in object representations. Supporting the view that multisensory advantages depend on specific conditions, semantically congruent audiovisual benefits may be restricted to memory recollection processes. Duarte et al. (2023) found improved recollection for visual objects previously paired with congruent sounds, even when sounds were task-irrelevant (Duarte et al., 2023). These findings suggest that object memory benefits from multisensory encoding, and that unattended auditory information can be semantically processed and integrated into memory representations. While the paradigm adopted by Duarte et al. (2023) was based on a delayed response, which in turn may have encouraged deeper memory retrieval processes (i.e.,

recollection), the paradigm adopted in Experiment 1B relies on faster-paced continuous recognition, where familiarity judgments may dominate instead. Moreover, in Duarte et al. (2023) the objects were visually rotated to match their natural orientation during sound production, possibly to enhance multisensory integration, while the stimuli presented in Experiment 1B were portrayed as ‘static’. Although cross-modal cues have been reported to influence early perceptual-stage on visual processing (Stoermer et al., 2009), findings from Duarte et al., (2023) work suggests that auditory cues might facilitate object discrimination particularly when presented closer to the decision-making phase of a task. This discrepancy suggests that the timing and relevance of sensory cues may critically modulate their effect on memory and recognition performance.

Nature of the auditory stimuli. The specific sounds adopted in the audiovisual pairings may not have fully captured the variability of object representations across categories (i.e., category-representative stimuli in both sensory modalities), limiting their effectiveness as multisensory cues. However, in Experiment 1A, we assessed the correspondence between object sounds and categories, and results indicated that most sounds were correctly matched to their respective basic-level object categories in a sound–object category matching task. The specific sounds adopted in the audiovisual pairings may not have fully captured the variability of object representations across categories (i.e., category-representative stimuli in both sensory modalities), limiting their effectiveness as multisensory cues. However, in Experiment 1A, we assessed the correspondence between object sounds and categories, and results indicated that most sounds were correctly matched to their respective basic-level object categories in a sound–object category matching task. This was true for the Bell sounds, but not for the Vase sounds, which was not correctly matched above chance. This is consistent with the partial ‘fit’ of the sound to match the variability of object representations in that specific object category. Thus, the sounds adopted to represent the Bell category were a better match than those related to the Vase category. The differential sound representativeness of object category led to different reaction times in the discrimination task of Experiment 1B: across all learning modalities, participants were

faster at discriminating Bell objects, suggesting a potential relationship between the sound representativeness of object category and responses time in visual objects discrimination. Moreover, in the sound-object exemplar matching task of Experiment 1A, we observed higher accuracy to object images than their word labels, indicating that sounds were more likely to match a specific category exemplar. This suggests that the audio-visual object pairings at the exemplar level might be a characteristic of object categories. However, when examining the results at the category level, Glass sounds were associated with increased accuracy when they were matched with corresponding words than visual object images. Our results might be due to individual variability (Hampton, 2006; Horowitz & Turan, 2008) in the 'representativeness' of the sounds or images across the object categories: participants might retain different prototypes representing the Glass category, for instance different sizes, and shapes of Glass. This finding alone suggests that familiar objects may not be suitable for investigating multisensory object categories and novel, unfamiliar objects might offer a more efficient way to investigate these processes, since they are not affected by prior experience or knowledge.

Finally, it is important to note that object sounds may be related to object shapes only in cases in which the objects are animate (Rubinstein, 2008), such as sounds generating from the function of the objects themselves (i.e. functional sounds), arising from the interaction with the objects (i.e. interaction-sounds), or, sounds that are related to the material properties of a specific object (i.e. material-sounds), and not to the object shape itself. Considering this distinction, the sound of the glass object could fit in the 'interaction' type of sounds (i.e., the clinking of a glass) whereas the sound of the bell may be related to its function. Thus, an 'interaction' or functional sound might activate a dynamic representation of the object, that in turn may be better represented by the object word (more abstract) and not by its static object image. We note that sound-object exemplar matching performance was below chance level for the Vase in Experiment 1A, suggesting the sounds did not represent the Vase object category, either at the basic or exemplar levels. In contrast, matching performance for the Bell sound was above chance for only the basic-

level category and not for the exemplar ones, indicating that Bell-sounds were representative of the basic-level category, but not of the specific object exemplars. Since Bell-sounds are inherently related to the material of the object itself (i.e., the sound produced by a heavy metal church bell), and the Vase-sounds generated for the experiment were 'material' and linked to the object material that might be produced, they may not have been as representative of the semantic associations in the representation of the object.

The absence of evidence for categorical perception (CP)

Results from Experiment 1B suggest no evidence for categorical perception (CP), both when the objects have been previously presented visual-only and paired with the semantic congruent sounds. Evidence for CP has been found for many levels of objects processing, across different domains, and at varying levels of stimulus complexity. For instance, CP has been reported both in one-dimensional and low-level visual discriminations for colour (Briscoe, 2020; Witzel & Gegenfurtner, 2016), spoken words (Simanova et al., 2016), phonemes (Kronrod et al., 2016); and in high-level visual discrimination for familiar objects (Newell & Bühlhoff, 2002), familiar faces (Beale & Keil, 1995), sex of faces (Bühlhoff & Newell, 2004) and facial expressions (Fugate, 2013). Evidence for CP has also been found in the tactile domain (Gaißert et al., 2012). Moreover, it has also been proposed that category selectivity arises after the learning process has occurred (Goldstone & Hendrickson, 2010): after the visual system develops expertise in distinguishing between highly similar but subtly different objects, then CP becomes clear. To illustrate, evidence of CP is shown for familiar and not for unfamiliar faces (Beale & Keil, 1995), and it can arise from learning unfamiliar faces (Goldstone et al., 2001). Despite this compelling evidence, overall, we failed to replicate the hallmarks of CP in Experiment 1B. In other words, participants did not show increased discriminability at the category boundary, even though they categorised objects into distinct categories, especially for within-object class. An explanation for these findings may arise from a difference in methodological factors related to the individual object continua (i.e., the morphing procedure), the nature of online testing; or to the sounds not representing the

individual object categories and possibly hindering perceptual performance. Regarding the morphed continua, the results from Experiment 1C, in which we investigated the perceptual similarity of morphed continua, suggested consistent similarity ratings across steps and categories, with smaller distances between objects on the morph continuum resulting in higher perceived similarity compared to larger distances.

Furthermore, the CP framework, both in speech and object perception research, proposes a discrete and distinct organisation of object information into perceptual categories. However, evidence from neural and behavioural research suggest that object categories are flexible and adaptable, instead of being isolated, as supported by feature-based processing (Kravitz & Behrmann, 2011; Op de Beeck et al. 2008), object constancy mechanisms (Vuilleumier et al., 2002), and contextual effects (Joubert et al., 2007). Importantly, McMurray (2022) challenges the traditional interpretation of categorical perception as a domain- and species-general phenomenon, by suggesting that what appears to be categorical perception may actually reflect task-specific effects rather than true perceptual categorisation. Specifically, McMurray argues that discrimination tasks do not necessitate categorisation, and they are characterised by high memory load involving encoding and retention processes. According to this view, the discrimination accuracy peaks, one of the hallmarks of CP, may reflect categorical memory or judgements instead of categorical perception (McMurray, 2022). Through his review, McMurray supports the view of continuous information organisation for categorisation, challenging the traditional CP framework that proposes a more discrete and distinct organisation of object information into perceptual categories. This perspective is supported by findings from several face and colour CP studies. Hanley and Roberson (2011) demonstrated the role of stimulus presentation order in discrimination tasks: when taking this factor into consideration, the results were best explained by a model of continuous encoding of stimuli, where categories could also influence discrimination accuracy. In another investigation on colour perception, Roberson et al. (2009) found that a new discrimination task, in which colours are present one after the other subsequently, led to the

absence of the CP effect. Furthermore, Van Geert & Wagemans (2023) introduced an important methodological consideration regarding the role of reference points in influencing categorisation and discrimination performance. Defining reference points as exemplars serving as comparison points, the authors found across different tasks - categorisation, discrimination, and similarity judgement - that category boundary (CP) effects are dependent both on the type of stimulus pair (whether within or between categories) and on the strength of individual reference points. Furthermore, the authors stressed the importance of using large sample sizes to allow for a nuanced understanding of individual and global performance.

According to McMurray's view, the lack of CP found in Experiment 1B might reflect continuous information organisation for categorisation, instead of a more discrete information organisation. This interpretation suggests that object perception might depend more on hierarchical information processing (Riesenhuber & Poggio, 1999), focusing on adaptive object representations that are dynamically updated through predictive processing mechanisms (Peelen et al., 2023; Kourtzi & Connor, 2011), rather than rigid categorical boundaries. Thus, CP might arise under specific experimental conditions or task demands, consistent with McMurray's argument that apparent categorical effects may reflect methodological artifacts rather than fundamental perceptual mechanisms.

2.9 Conclusion

Chapter 2 investigated whether audio–visual exposure influences the visual categorisation of familiar objects and assessed the representativeness of auditory stimuli for their corresponding object categories. Overall, we observed no multisensory benefit on the category boundary in categorising familiar objects, nor evidence of categorical perception (Experiment 1B). Although we found no evidence for categorical perception across either unimodal or audio–visual learning conditions, participants perceived the objects as belonging to distinct categories, particularly for within-category items, even in the absence of increased discriminability at the category boundary.

Sounds were generally matched to their corresponding object categories at above-chance levels (Experiment 1A), and the morphed continua were rated as consistently similar across steps and categories (Experiment 1C). The lack of evidence for categorical perception may reflect continuous object representations during categorisation tasks, individual differences in object representations, or methodological differences relative to previous paradigms (e.g., stimulus set, task demands, or paradigm structure). Our findings do not exclude that object representations may still be based on a combination of relevant information from multiple sensory inputs.

Chapter 3

The way an object moves, but not its sound,
influences its perceived similarity to other object
shape

3 The Way an Object Moves, but not its Sound, Influences its Perceived Similarity to Other Object Shape

3.1 Abstract

Perceived shape similarity between objects is central to many cognitive models of object categorisation and there is growing evidence that shape similarity acts as an organisation principle in the neural representation of objects. However, many objects are associated with distinct patterns of motion and sounds, and it is unclear how these features influence the manner in which objects are represented. For example, a tennis ball has a shape that is very similar to other balls, such as foot- or golf balls, but its movement when hit, or its bouncing sound are more distinct. The similarity of an object to other objects may be determined by an integrated representation of its shape, movement and sound or by each dimension separately. We investigated whether motion and sound directly influence the perceived shape of novel objects using judgements of inter-item similarity. In Experiment 1, we found that object shapes which shared similar motion patterns were perceived as more similar than shapes with inconsistent or no motion. However, the addition of (movement-generated) sounds did not further influence judgements of shape similarity. In Experiment 2 we employed a Garner interference task to determine how shape and motion are processed in object perception. The results suggest that object motion is integrated into the representation of object shape in memory. Our findings have important implications on our understanding of how inter-object similarity is determined from cue combination and the consequent role of shape and motion features in the perception and categorisation of objects.

3.2 Introduction

In the real world, objects are categorised and recognised on the basis of multiple featural dimensions including spatial (Marr & Nishihara, 1978; Biederman, 1987; Tarr & Bülthoff 1995), motion (e.g. Scholl, 2007; Johansson, 1973), auditory (e.g. Molholm et al., 2004; Giard & Peronnet, 1999) or even tactile (e.g., Klatzky et al., 1985; Lederman & Klatzky, 1987) dimensions. The range of multidimensional, and multisensory, object features that represent objects can be large, which underpins the complexity involved in understanding how these representations are structured in the brain to allow for recognition as well as generalisation to novel exemplars. Evidence from behavioural (Rosch et al., 1976a; Goldstone, 1994a; Cutzu & Edelman, 1998; Smith & Heise, 1992) and neural (Haxby et al., 2001; Grossman et al., 2002; de Beeck et al., 2008; Cichy et al. 2019) investigations of object perception suggests that perceptual similarity is an important guiding principle for organising sensory information into object categories. Indeed, influential theories based on inter-object similarity (Nosofsky, 1986) or ‘family resemblances’ (Rosch & Mervis, 1975; Rosch et al., 1976b) have been supported by evidence from a large number of empirical studies (e.g. Goldstone et al., 2001; Newell & Bülthoff, 2002; Mervis & Rosch, 1981; Medin et al., 1987; Smith & Minda, 2002; Wu et al., 2020). Collectively these studies suggest that an object’s representation in memory depends on its psychological distance in representational space from other representations, with inter-item similarity based on the number of shared features or dimensions with other exemplars or relative to a prototype. However, given the complexity of object representations, it is unclear how featural similarity applies to information from different dimensions, including across modalities, to result in a coherent percept or category. To investigate this, we conducted two experiments in which we tested whether the perceived inter-object similarity of novel object shapes was affected by the object’s motion or object sounds.

An important aspect of theories based on ‘object’ or shape spaces is that these spaces are necessarily flexible, and their structure can depend on available sensory information or top-down

goals. For example, some features may be more perceptually salient than others, or more frequent, and therefore may dominate the structure of the representational space (or provide the scaffold to which other features are integrated, e.g., Goldstone, 1993; 1994b; Goldstone et al., 2001; Folstein et al., 2013; Carvahlo et al., 2021). Take, for example, the difference between the shape and movement of different types of felines: shape or sound are more salient than motion for discriminating between tigers and domestic cats. Alternatively, features may be reweighted by cognitive factors including attentional focus (Nosofsky, 1986; Love et al., 2004), learning (Goldstone et al., 2001), adaptation (Leopold et al., 2001), or task demands (Goldstone, 1994a; 1994b; Kruschke, 1993; Deng & Sloutsky, 2015), affecting their relative similarity or discriminability (e.g. Leopold et al., 2001). Thus, as you walk through a jungle in India, it may be more expedient to perceive a cat's movement as it approaches than either its shape or sound.

Theories of object categorisation that can account for this flexibility typically assume that the underlying dimensions of the representational space are separable (Garner & Felfoldy, 1970; Ashby & Townsend, 1986; Ashby & Maddox, 1994; Poggio & Girosi, 2002). However, some studies of object perception suggest that multidimensional or multisensory features may be non-separable or fully integrated. In support of this, a number of studies on multisensory feature integration have, for example, reported a greater benefit on perception from a combination of crossmodal inputs than is predicted from performance based on each feature alone (Murray et al., 2005; Molholm et al., 2004; Kyler & James, 2024), known as 'superadditive' performance (e.g. Meredith & Stein, 1983). The findings of several studies of object motion have suggested that movement is integrated into the representation of an object's shape, with a resulting benefit on the recognition (Chuang et al., 2012; Stone, 1998) and categorisation (Newell et al., 2004; Setti & Newell, 2010) of novel objects (see also Liu & Cooper, 2003; Vuong et al., 2009; Wallraven & Bülthoff, 2009). In a similar manner, other studies have reported that semantically related sounds benefit the recognition of familiar objects (Chen & Spence, 2011; Mädebach et al., 2017); also suggesting that objects shapes and sounds are integrated (Mulatti et al., 2014). In further support

of integrated dimensions, visual images are more memorable if paired with distinctive sounds during learning than more typical sounds (Bülthoff & Newell, 2015). Finally, object sounds can enhance visual motion when both are correlated (Gleiss & Kayser, 2014) or meaningfully related (Thomas & Shiffrar, 2013), suggesting that sound and motion may also be integrated in certain circumstances.

When a new dimension is introduced during category learning, the brain reconfigures the relevance of each dimension, a process known as dimensionality modulation (Folstein et al 2013; Doshier & Lu, 1998). This modulation is a form of perceptual restructuring (or ‘stretching’) making category-relevant dimension more salient and reduces (or compresses) the ones that are category-irrelevant. For example, if categories were initially defined by a shape, adding a new predictive colour dimension can cause a perceptual shift where colour becomes the primary cue, and the previously important shape information is adjusted to become a secondary cue. In support of this, studies on category learning have, for example, reported enhanced discrimination along category-relevant dimensions, while worsening it along irrelevant ones (Goldstone, 1994b; Gureckis & Goldstone, 2008; Folstein et al., 2012; 2013). This benefit appears to selectively stretch dimensions that are either global to the relevant dimension (Goldstone, 1994; Op de Beeck et al., 2003) or limited around the category boundary (Goldstone et al., 1996; Notman et al., 2005; Özgen & Davies, 2002). This dimensional reweighting appears to be an ‘error-driven’ process, guided by knowledge and feedback, where the restructuring and reweighting of features improve performance with new information (Kalish et al., 2005; Sewell & Lewandowsky, 2011). Neuroimaging studies provide evidence for neural plasticity in the visual cortex due to visual category learning (Kourtzi et al., 2005; Folstein et al., 2013; 2015; Pèrez-Gay Juárez et al., 2019) expertise to new categories (Gauthier et al., 1999; 2000), and these changes can be specific, enhancing the representations of exemplars near to the category boundary (O'Bryan et al., 2024; Henderson et al., 2025).

Object shape is often considered separable in categorisation tasks requiring fine-grained discrimination (Garner, 1974; Shepard, 1964; Goldstone & Steyvers, 2001), whereas motion may act as an integral dimension, especially for unfamiliar or visually similar stimuli, where its co-variation with shape helps resolve ambiguity (Ullman, 1979; Grzywacz & Hildreth, 1987), support perception (Stone, 1998, 1999; Chuang et al., 2012) and guide category membership decisions (Edelman, 1998; Tjan & Legge, 1998; Vuong & Tarr, 2004; Vuong et al., 2005). Understanding whether these features are processed integrally or separately is critical for understanding how perceptual similarity informs the representations underpinning object categories. In particular, investigating how motion and sound congruency influences the perceived similarity of object shapes could offer a better understanding of the dimensional architecture of object representations. If feature dimensions are separable, we might predict, on the one hand, that object motion or its sound may not affect an object's perceived similarity to other objects and shape is a dominant feature. Alternatively, object motion and/or its sound may be more salient features than object shape and therefore have a greater influence on inter-object similarity than its shape. In that case, objects that move or sound the same may be perceived as more similar irrespective of their shape (i.e. whether shape is the same or different). Thus, this would suggest that separable dimensions can be reweighted, making a more salient dimension the basis for similarity judgments. If features are integrated, they might be processed as a single unit combined into a unified percept, then we might expect judgements of the similarity of the same object shapes to increase when their motion (or sound) is the same relative to judgements of the same shape only. However, any differences in motion or sound would not further influence similarity judgements if the shapes were different.

3.3 Outline of the Experiments

Across a series of experiments, which were conducted online, we examined how object motion influences the perceived similarity of the shapes of novel objects (Experiment 2A), and

whether the addition of correlated sound further influences these judgements (Experiment 2B). A central issue is whether object shape and its motion are represented as separable or integral dimensions. Therefore, in Experiment 3, we adapted the Garner interference paradigm, in which participants classified stimuli based on one relevant dimension while ignoring an irrelevant dimension, to directly test whether shape and motion act as integral or separable dimensions in object categorisation. According to Garner's (1974) dimensional framework, separable dimensions can be attended to independently, while integral dimensions elicit interference when varied unpredictably.

3.4 General Stimuli

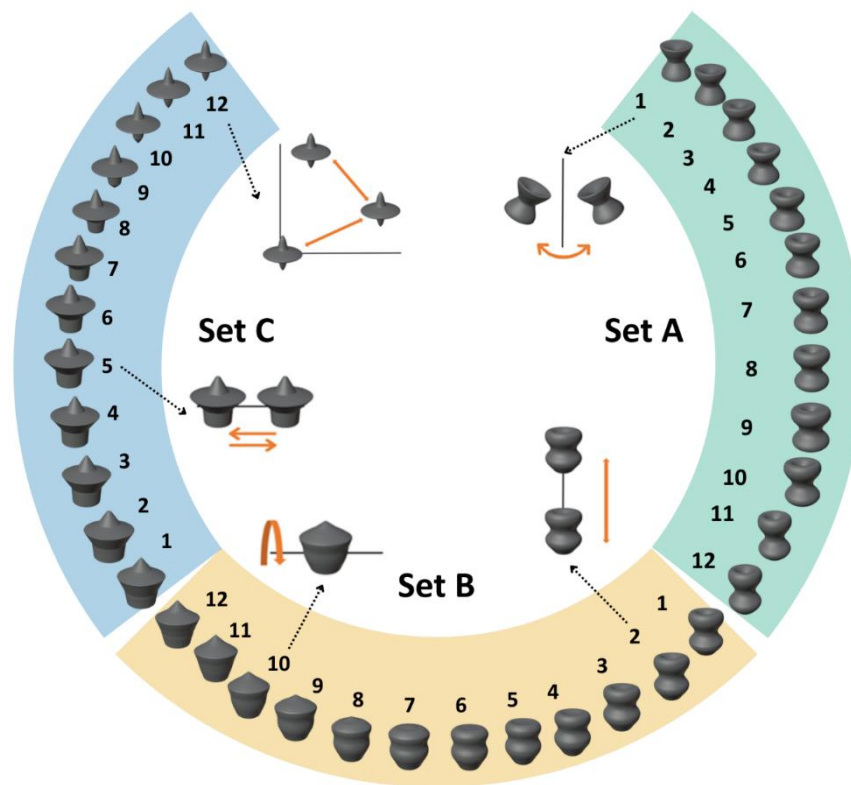
We selected 36 novel object shapes from shape space adapted from previous studies (Li et al., 2020; O'Dowd et al., 2025). Within the shape space, 3D object shapes were created utilizing 3D modelling software (Blender Foundation, 3.5.0, 2023, www.blender.org), and designed to be 10 degrees apart from their neighbours, except for objects positioned the end points of the space (due to the curvature introduced during 3D rendering). The object creation followed a specific pipeline: (1) each shape was converted from a .png image to a Scalable Vector Graphics (.svg) format using www.pngtosvg.com/, maintaining three constant colour levels (e.g., white, black, grey); (2) each vector was then imported into the 3D environment; (3) the outline of the imported vector was isolated and converted into a mesh; (4) this mesh was rotated 360 degrees around its central vertical axis using Blender's Spin Function in Edit mode (360 degrees, 100 steps); and (5) each complete 3D shape was extracted. This pipeline was consistently applied to generate all object shapes used in the experiment. The 3D space, lighting (point light, 0.1m radius, 1000W; coordinates: 11m, -14m, 6.9m; rotation: 40°, 34.8°), and viewpoint (coordinates: 10.9m, -14m, 4m; rotation: 70°, 0°) settings were standardized across all shapes. Each 3D object was rendered using the Workbench Engine (32 render samples, Single pass Anti-Aliasing viewpoint; Studio Lighting, dark grey Colour Material (RGB: 107, 109, 109, 254; HEX: #6b6d6d), and Specular

Lighting). All object images were extracted at a resolution of 1080 x 1080 pixels, 100% scale, and presented in a canonical, 3/4 view to ensure the 3D object and its pertinent features (e.g., concavities) were visually discernible. The shape space depicting these objects is shown in Figure 16 and the individual 3D objects utilized as stimuli are accessible on the Open Science Framework page for Experiment 2: <https://osf.io/s369c/>.

We then animated each of the 36 rendered 3D objects according to one of five distinct motion patterns (each motion was randomly assigned to each object, see Design section): swing, jump, roll, or shake abrupt movement (see Figure 16). The selection of these motion patterns was based on their movement characteristics (e.g., smooth or abrupt) and their reference axis (e.g., horizontal or vertical movement of the object). Each motion animation was rendered using the same Workbench Engine settings as described previously. All five object movements had a consistent duration of 2 seconds. Each moving object was displayed against a background created in Blender, consisting of two grey walls and a grey floor. The camera angle (coordinates: 10.9m, -14m, 4m; rotation: 70°, 0°) and lighting (point light, 0.1m radius, 1000W; coordinates: 11m, -14m, 6.9m; rotation: 40°, 34.8°) remained constant. The light source illuminated the scene from above, casting a shadow from the object onto the floor to enhance the perception of depth and motion.

Figure 16

An illustration of the 'shape space' of object stimuli used in the Experiments 2 and 3. Objects were assigned to one of 3 sets (A, B or C). Motion condition (static, consistent or inconsistent) was randomly assigned each object set (e.g. objects in Set A were all presented with the consistent motion).



Note: The shape space was adapted from a previous circular shape space described by Li et al., (2020) and O'Dowd et al. (2025). The objects are arranged within the shape space such that neighbouring objects are equal in angular distance from each other. Within the 'shape space' are illustrated example of the five motion patterns associated with the objects (from left to right): abrupt motion, shake, roll, jump, swing.

In addition to the five object movements, five sound clips were also used as stimuli in our study (see Experiment 2B). Each sound was correlated with its respective movement. To generate these sounds, we first downloaded individual sounds from online sound databases (e.g., Free

Sounds Library and freeSFX). The sound clips selection was based on their perceptual resemblance to the intended motion patterns (e.g., swing, jump, roll, or shake abrupt movement). The sounds were validated in an online pilot study using a two-alternative forced choice (2AFC) discrimination task: participants were asked to match each sound clip to one of two motion patterns presented. This allowed us to test whether the sound clips were reliably discriminable as representative of their associated visual motion patterns. All sound clips were matched to the relative visual motion pattern well above chance, with mean accuracies ranging from 0.75 to 1.00, confirming that participants were able to reliably associate the sounds with their corresponding motion patterns (for more details of the sound validation study refer to Supplementary Materials – S8). The sound stimuli for the experiment were reproduced using the software Audacity (Audacity Team, 2023, v3.4.2). The sound clips were generated using the same tone (B4), to ensure pitch consistency across stimuli. All clips were normalized to a uniform peak amplitude and corrected for DC offset (a constant shift of the waveform above or below zero on the amplitude axis) to avoid introducing unintended artifacts. Each sound designed to correlate and fully synchronise with a corresponding object movement pattern. We achieved this in two steps as follows: first, we identified the keyframes that defined each object movement pattern (e.g., the extreme positions of a swing at the far left and far right of its arc). Second, we mapped these keyframes into the sound temporal structure by aligning changes in amplitude or modulation with the corresponding motion points. For example, in the swinging motion, the sound would peak when the object reached the turning points of its arc, thus replicating the rhythm of the visual swing. All five sound clips had a duration of 2 seconds (to match the object motion) and a rate of 44100 Hz, meaning the audio waveform was digitised at 44,100 samples per second to ensure high-quality playback across devices.

The display of each object stimulus in a trial was masked. We created individual object masks by scrambling the image of each object shape using a MATLAB script (MathWorks, 2023). The script performs a controlled scrambling of the central region of each individual image. To achieve this, it divides the original image on the object into 12x12 pixel tiles, then scrambles the

central tiles, and subsequently extracts a central region of 24x24 pixels from the scrambled image. Finally, this central region is reconstructed into tiles, which are then randomly shuffled. (See the Project page on the OSF repository for the MATLAB script).

The experiments were developed in PsychoPy (Pierce et al., 2019, 2022, v2023.2.3), delivered online through Pavlovia (<https://pavlovia.org/>), and it was constrained via Pavlovia to run only on desktop computers.

3.5 Experiment 2A

Experiment 2A investigated how visual motion influences the perceived similarity of the shape of novel object pairs, building on evidence that coherent movement can promote perceptual grouping (Wertheimer, 1938; Johansson, 1973; O'regan et al., 1999; Murray et al., 2005). This experiment investigated whether objects moving in a consistent manner would be judged as more similar than objects with inconsistent or static motion. In this experiment, participants rated the shape similarity of object pairs across three object motion conditions (static, consistent movement, and inconsistent movement). Importantly, participants were instructed to base their ratings on object shape alone, ignoring any other information.

We use nonmetric multidimensional scaling (NMDS) technique to reconstruct stimuli space across the motion conditions. Multidimensional scaling (MDS) provides a means to map perceptual similarity, projecting similarity judgments into geometric space to reveal underlying perceptual dimensions (Kruskal & Wish, 1978). Prior work shows integration of shape and motion in object representation (Cutting & Kozlowski, 1977) and that motion contributes to perceptual organisation by segmenting scenes into perceptual units (Scholl, 2001). Dynamic cues, such as characteristic object motion, influence category formation (Newell et al., 2004; Vuong & Tarr, 2006) and, as static features of shape or texture, support defining category structure (Tarr & Gauthier, 1998; Barsalou, 1999). Motion is also encoded in long-term memory and supports categorisation (Konkle & Oliva, 2012), underlying its role in similarity judgments and perceptual

grouping. Multisensory studies indicate that similarity judgments combines and weights information across modalities (Spence & Driver, 2004), rather than treating the inputs in isolation. For instance, a judgment about an object relies on integration of both its visual and auditory properties. This integration is governed by the spatial and temporal proximity of the cues; stimuli that occur at the same time and place are more likely to be bound together into a single object or event (Welch & Warren, 1980; Wallace et al., 1998). Taken together, these findings suggest that consistent motion might enhance perceived similarity between objects, particularly for identical shapes, while inconsistent motion may disrupt such grouping.

In this study, the static condition served as a control condition in which similarity judgments would reflect uniquely to object shape. We predicted three possible patterns: (i) any motion, regardless of type, might increase perceived similarity relative to the static baseline (consistent = inconsistent > static); (ii) only consistent motion might enhance similarity, while inconsistent motion would resemble the static condition (consistent > inconsistent = static); or (iii) motion might differentially affect similarity such that consistent motion increased, and inconsistent motion reduced, perceived similarity (consistent > static > inconsistent). Finally, it was anticipated that motion might affect ratings mainly for same object pairings and may have no or reduced effect when object pairings differed in shape.

3.5.1 Method

Participants

We recruited 45 adult participants (mean age = 39.5 years, SD = 11.4; 48.9% female) for our experiment using Prolific (<https://www.prolific.co/>) the majority of whom completed at least second-level education (94.4%; 80% completed third level education). Our inclusion criteria were as follows: fluency in English, normal or corrected-to-normal vision and no hearing impairments. All the participants were naïve to the purpose of the study and received compensation to take part in the experiment (at a rate of £9.00 per hour). The study was approved by the Trinity College

Dublin, School of Psychology Research Ethics Committee (approval number SPREC102020-50) and complied with EU data protection legislation (GDPR).

Design

The experiment was based a two-way, within-subjects design, with motion condition and object pairings (same or different according to intervening steps) as the main factors. There were 3 motion levels: no object movement (i.e. static objects), consistent object movement, and inconsistent object movements. The object pairs were either the same object (0 steps) or differed between 1 (e.g. object 1 and 2) and 11 steps (i.e. object 1 and 12). The main dependent variable was similarity ratings across the pairs of objects.

For the purpose of our experiment, particularly to ensure that motion was fully counterbalanced across object shapes, we first divided the 36 object stimuli into three sub-sets of 12 object stimuli each (see Figure 16). For any one participant, each motion condition (static, consistent or inconsistent) was then randomly assigned to one of these object sets (e.g. objects in Set A were all presented with the consistent motion). This assignment of motion condition to objects within each object set was fully counterbalanced across participants. Each object was compared once with itself and once with every other object within its set, leading to 78 trials per object set (i.e. $12 + (12 \times 11)/2$). Trials containing object stimuli from within each set were presented in blocks and block order was counterbalanced across participants. We decided to block the trials according to the motion condition to ensure that participants were judging the similarity of the objects shape only: an interleaved trial design may have made this task more challenging (Carvahlo & Goldstone, 2014a;2014b; 2015). Trial order within each block was fully randomised. Overall, each participant was presented with 234 (78×3) trials representing all 36 objects in the shape space.

Procedure

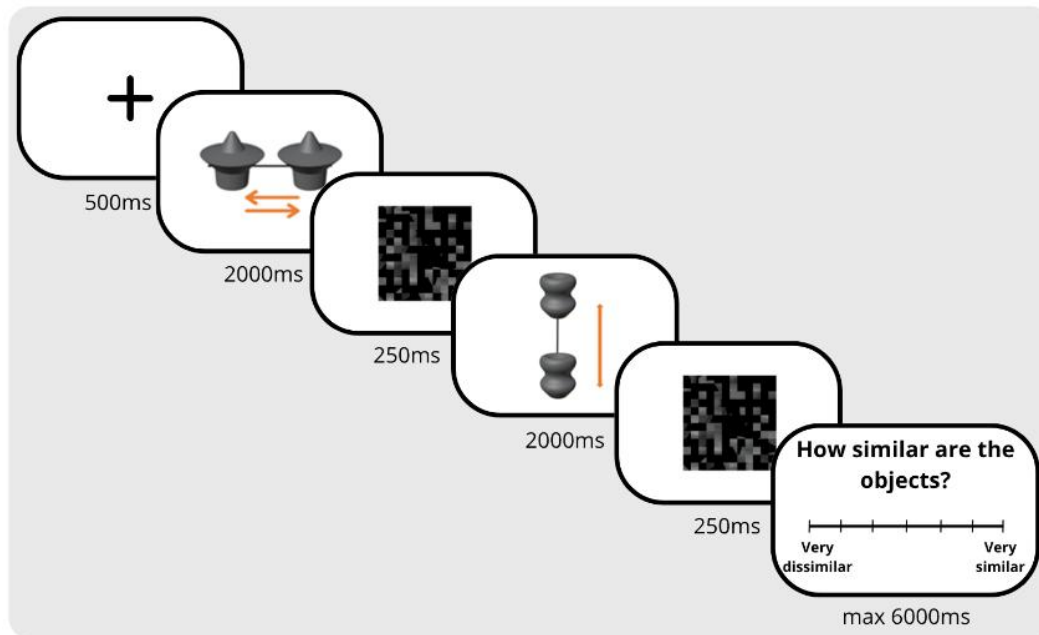
Prior to the experiment, all participants provided informed consent to proceed. This was then followed by a basic questionnaire on demographic information. The task instructions

explained that their task was to examine pairs of object stimuli presented in each trial and to rate their similarity by moving their computer mouse over a sliding scale which represented a 7-point Likert scale, ranging from 'very dissimilar' (1) to 'very similar' (7). They were further instructed to base their similarity judgements only on the shape of the object and to ignore any other object information, such as its motion. Participants were informed to use the whole scale to express their judgements.

The experiment began with a short block of practice trials, to allow participants to familiarise themselves with the task, which was then followed by the main task. A schematic illustration of the sequence of events in a typical trial is presented in Figure 17. In each trial, a fixation cross appeared on the centre of the screen for 500ms, after which the object stimulus was presented at the centre of the screen. The object may or may not be presented as moving (depending on the experimental condition). The object stimulus was presented for 2 seconds and then followed by the corresponding visual mask for 250ms. Then the second object stimulus was then presented (at the centre of the screen) for 2 seconds and again followed by the corresponding mask for 250ms. After the second mask disappeared, participants were presented with the response screen in which they were asked to provide their rating of how similar the two successive objects were (from 1 to 7). They were given a maximum of 6 seconds to make a response after which the trial ended, and no rating was recorded. Participants were provided with feedback on the timing of their response if they exceeded 2 seconds: in that case they were presented with *'Too slow! Please respond faster'*. At the end of the experiment, participants were asked to provide a brief description of how they had judged similarity among the objects. Participants took approximately 30 minutes to complete the whole experiment.

Figure 17

A schematic illustration of the timeline of a typical trial in Experiment 2 (from left to right).



Note: Each trial started with a fixation cross (left), followed by an object stimulus (here depicted in the 'vibration' movement), an object mask, followed by the second object stimulus (here shown the 'jump' movement) again followed by a mask. A response screen was provided to allow the participant to provide their similarity rating.

Data Analysis

Differences in similarity ratings across the 3 stimulus sets for the static condition only were first assessed to ensure that assumptions for parametric testing were met. To investigate whether shape similarity judgments were affected by object motion, we normalised the participants' ratings by transforming each individual data point (e.g., similarity ratings) to a range between 0 and 1 using the Min-Max Normalisation function (Patro & Sahu, 2015). We labelled each of the 12 objects within the 3 object sets (see Design) according to their order in the shape space, such as if they were in the first, second, third, position these were labelled 1, 2 and 3 respectively. Thus, we used 12 labels associated with the objects. This approach was adopted for both the analysis of similarity matrices and for multidimensional scaling. After data normalisation, the pairwise similarity ratings were then averaged across all participants for all trials, in order to generate

similarity matrices for each motion condition separately. This allowed us to develop corresponding 'heatmaps' to illustrate the perceived similarity, with lower similarity ratings defined by light yellow colour and the higher similarity ratings by a red colour. For both similarity ratings and reaction times (RTs), we subdivided the dataset according to the identity of the object pairings within a trial, i.e. whether the object pair was the same object (e.g., object 1 compared to 1, or 2 compared to 2) or different objects (e.g., object 1 compared to 3 etc.) and the relative step difference between the different objects (e.g. 2 step difference between objects 1 and 3). We then analysed this dataset by fitting a linear mixed-effects regression model to the data, considering the motion conditions (consistent, inconsistent or static) and object pairings (same or different objects) as factors.

We also analysed the similarity ratings using a nonmetric multidimensional scaling (NMDS) technique to explore the psychological structure in our data set (Borg et al., 2012) of object shapes as a consequence of object motion: MDS helps identify the configuration of object coordinates in a low-dimensional space which preserves the pairwise dissimilarities/distances between objects. Nonmetric MDS (NMDS) requires as input dissimilarity/distance matrices, and it returns as output a two column vectors with coordinates for each stimulus in the output space, preserving the rank order of the input dissimilarities. Although, similarity data obtained from Likert scales are not directly considered as 'distances', the correlation coefficients among multiple columns do. Thus, to calculate the dissimilarity matrices for the objects as a consequence of each motion condition, we computed Pearson's r and the corresponding correlation matrices of the normalised data. The correlation matrices were then transformed in dissimilarity matrices by subtracting the Pearson's r from 1 ($\delta = 1 - r$). Once we obtained the dissimilarity matrices, we then transformed them into distance matrices (Euclidean distances), defined as mirrored matrices along their diagonal, which were generated for the object pairs in each motion condition. We then evaluated stress functions with dissimilarity matrices as input, to determine the optimal number of dimensions for the NMDS solution. To that end, we generated a scree plot from the MDS

analysis for objects in each motion condition, these plots graph the stress function value (y-axis) against the number of dimensions (x-axis), allowing us to determine the optimal dimensionality for our data. The stress function, through Kruskal's stress formula 1 (Kruskal, 1964a, 1964b), is a fit between input proximities and distances and it was adopted to determine the adequate number of dimensions to represent the output space reconstruction. Typically, a lower stress value indicates a better fit: stress values below 0.2 are an indication that the output space dimensionality is a good estimation of the input distances (see e.g. Cooke et al., 2006; Cox, 2001). We then performed nonmetric MDS analysis on these dissimilarity matrices and reconstructed the 2-dimensional shape space for each motion condition. The NMDS output configurations were then investigated and interpreted by visual inspection. To further analyse the clustering of the stimuli, we applied the Calinski-Harabasz index to determine the optimal number of clusters. The Calinski-Harabasz index, a method proposed by Calinski and Harabasz (1974), estimates the quality of the clustering by considering the ratio of the sum of between-cluster dispersion to within-cluster dispersion.

The data in all experiments were analysed with R via RStudio (R Core Team, 2021). The MDS algorithm used in Experiment 1 was from the 'smacof' package (Mair et al., 2022). Models were fitted using the 'lme4' package (Bates et al., 2014). Post-hoc comparisons were conducted using the 'emmeans' package (Lenth Russell et al., 2022) and the Bonferroni correction was applied to correct for multiple comparisons. Experiment 1a was pre-registered prior to data analysis on the Open Science Framework, <https://osf.io/raksy>, and further details, including the analysis script, are also available at https://osf.io/aquxv/?view_only=68d7402177714072aa0b220e609ecfdf.

3.5.2 Result

All participants took part in the experiment on desktop computer and accessed the experiment using different web browsers (84.4% used Chrome, 11.1% Edge and 4.5% used

Firefox). Participant data was screened for quality including usage of the rating scale and time taken to respond to each trial (with outliers defined as ± 3 SD away from the group mean or ± 4 SD away from the individual mean, consistent with previous reports e.g. Williams et al., 2022). All participants displayed evidence of using the whole rating scale in their responses. Within these response times, outliers were excluded from the data prior to analysis and represented of 5.3% of the entire RT data set. As an exploratory analysis to better understand participants self-reported categorisation strategy, all participants provided a free-typed description, post-experiment of how they conducted the task. The words 'shape', 'shapes' or 'object' were the most frequently reported, whilst other spatial words (size, curvy, sharp, edges, width) were less frequently mentioned. There was only one mention of object motion in these reports. Further details can be found under Supplementary Materials – Figure S7A.

Similarity rating data

Table 4

Mean normalised similarity ratings between pairs of objects across the step sizes and motion conditions in Experiment 2A.

Motion	Step size											
	0	1	2	3	4	5	6	7	8	9	10	11
Static	0.88	0.77	0.65	0.57	0.47	0.42	0.35	0.33	0.29	0.24	0.25	0.19
	0.20	0.26	0.29	0.30	0.30	0.31	0.29	0.30	0.30	0.27	0.27	0.26
Consistent	0.92	0.79	0.71	0.62	0.52	0.45	0.39	0.32	0.29	0.26	0.25	0.30
	0.16	0.27	0.29	0.31	0.32	0.32	0.30	0.30	0.29	0.27	0.28	0.30
Inconsistent	0.84	0.73	0.64	0.54	0.46	0.38	0.34	0.30	0.26	0.21	0.19	0.19
	0.20	0.28	0.28	0.30	0.29	0.27	0.28	0.27	0.25	0.23	0.24	0.23

Note: Step size refers to the relative position of each object in the pair within the shape space

(e.g. referring to Figure 16, the object pairs might include object 1 and object 11 resulting in a

step size of 10). A step size of 0 means the objects were the same. Standard deviations are shown in italics.

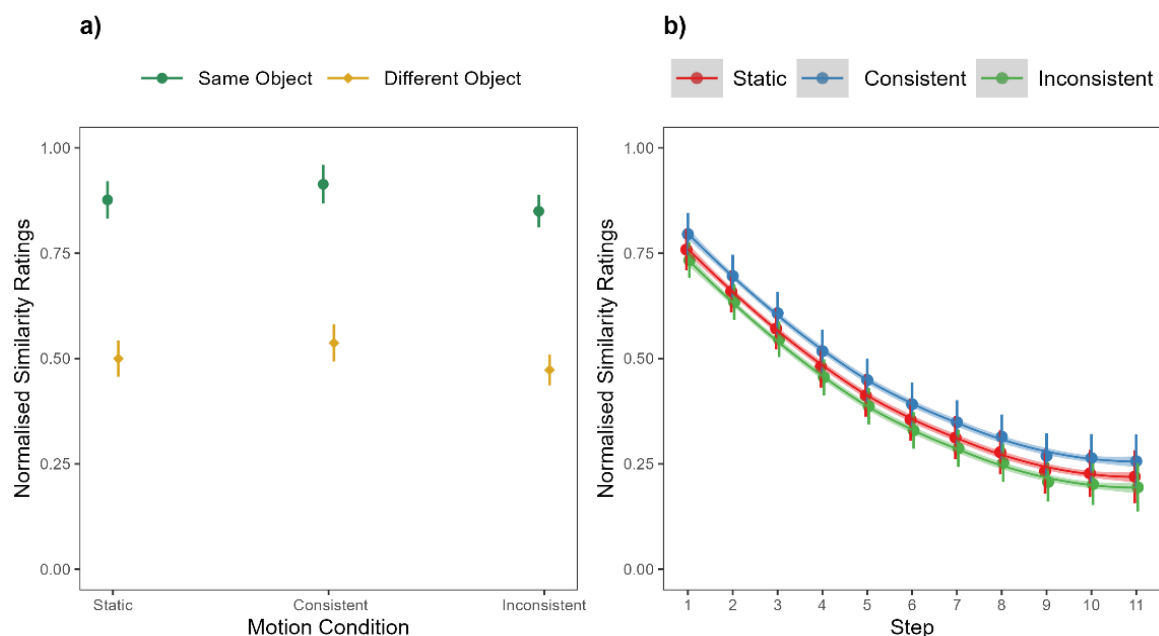
A comparison of ratings across the 3 object sets revealed that the assumptions for parametric testing were not met: a subsequent non-parametric Kruskal–Wallis test revealed a significant effect of stimulus set, $\chi^2(2) = 97.15$, $p < .001$. To account for this variability, stimulus set was included as a random intercept in the mixed-effects models. The mean similarity ratings for each of the motion (static, consistent or inconsistent) and object (same, different) conditions are presented in Table 2. As described above, we used a linear mixed-effects regression model to analyse the similarity ratings data and conducted a comparison between models with and without the interaction between motion condition and object pairing (same or different). A likelihood ratio test indicated that the interaction term did not significantly contribute to the model predicting similarity ratings, $\chi^2(2) = 0.38$, $p = .83$. The model without the interaction term showed significant main effects of both motion condition, $\chi^2(2) = 10.21$, $p = .006$ and object pairing, $\chi^2(1) = 2091$, $p < .001$ (see Figure 18A). Across conditions, pairings involving the *same object* were rated as significantly more similar than *different object* pairings (estimate marginal means, EMM = 0.38, 95% CI [0.34, 0.41], $p < .001$; see Figure 18A – yellow circles). Post hoc comparisons on the motion condition revealed significantly higher similarity ratings to object pairs with the *consistent movement* than object pairs with an *inconsistent movement* (EMM = 0.06, 95% CI [0.02, 0.11], $p = .003$). However, there was no difference in similarity ratings between no (i.e. *static*) movement compared to *consistent movement* (EMM = -0.04, 95% CI [-0.09, 0.02], $p = .36$) and inconsistent movement (EMM = 0.27, 95% CI [-0.03, 0.08], $p = .77$).

To investigate these effects in more detail, we then analysed the effect of motion condition separately for the *same object* and *different object* pairings across the intervening step sizes between the object pairs. For the same object pairings, a comparison between models with and without the motion condition revealed that motion significantly contributed to the model

predicting similarity ratings, $\chi^2(2) = 19.08$, $p < .001$. Post hoc comparisons showed that whilst objects with the *consistent movement* were rated as significantly more similar than those with *inconsistent movements* (EMM = 0.08, 95% CI [0.03, 0.13], $p < .001$), and there was no difference in similarity ratings to *static* object pairs relative to ratings in either the consistent or inconsistent motion conditions (see Figure 18A). For the different object pairings across the step sizes (see Figure 18B), the reduced model without the interaction showed significant main effects of both motion condition, $\chi^2(2) = 7.36$, $p = .02$, and step, $\chi^2(10) = 3693$, $p < .001$. Across step levels, object stimuli with the *consistent movement* were rated as significantly more similar than objects with *inconsistent movements* (EMM = 0.06, 95% CI [0.007, 0.12], $p = .023$), as shown in Figure 18B. Finally, there was no difference in similarity ratings to *static* object pairs relative to ratings in either the consistent motion condition (EMM = -0.04, 95% CI [-0.01, 0.03], $p = .55$), and the inconsistent motion condition (EMM = 0.03, 95% CI [-0.04, 0.09], $p = 1$).

Figure 18

The plots show the main results based on similarity ratings in Experiment 2A.



Note: Panel a) Main predictions from the linear mixed-effects models on the probability of similarity ratings (estimated marginal means) for the motion and object conditions. Error bars

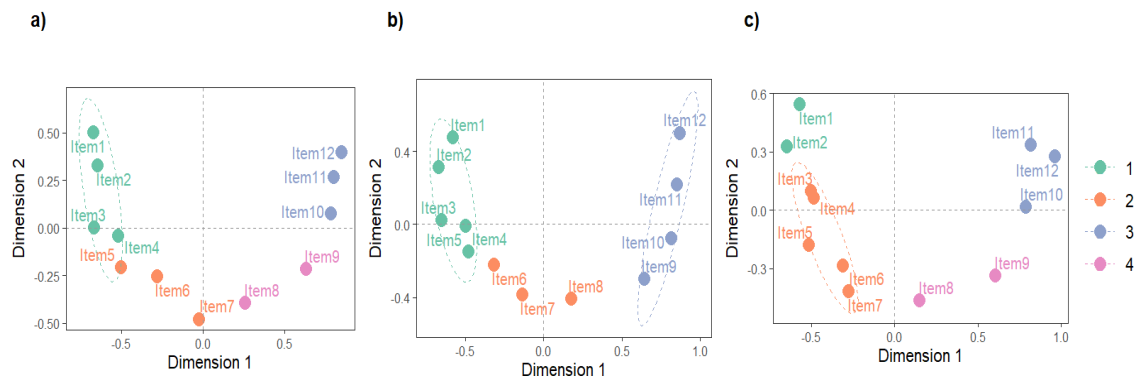
represent 95% confidence intervals. b) Mean similarity ratings to each of the different object pairings with increasing step size between objects. The solid line represents a weighted regression (*loess*) fit, and the shaded area indicates the 95% confidence interval around the smoothed trend.

Multidimensional Scaling

As detailed above, separate non-metric multidimensional scaling (NMDS) analyses were conducted on (dis)similarity matrices to each of the movement (static, consistent and inconsistent) conditions using ordinal-level constraints. Dimensionality was determined via scree plots and stress reduction curves, which indicated a two-dimensional solution was optimal across all conditions (Kruskal, 1964a). The 2D configurations provided good fit, with stress values of 0.012, 0.008 and 0.012 to the static, consistent and inconsistent movement conditions respectively (for further details refer to Supplemental Materials – S6). Figure 19 displays the relative position of the objects across each of the movement conditions. Dimension labels were interpreted by visual inspection of the output configuration. Each point in the plot corresponds to a stimulus, with the distances between points reflecting the perceived dissimilarities. K-means clustering was applied to the MDS coordinates which identified 4 clusters in the Static and Inconsistent movement conditions, and 3 clusters in the Consistent movement condition, consistent with Calinski–Harabasz index results. The resulting configurations (Figure 19) illustrate that consistent motion conditions led to more compact and separable perceptual clusters compared to static condition.

Figure 19

The plots show the two-dimensional non-metric MDS solutions in Experiment 2A.



Note: K-means clustering results overlaid, for the object similarity ratings when objects pairs had a) no movement (i.e. static), b) the consistent movement or c) inconsistent movement. The position of each point represents their perceived similarity to other items within that condition and dotted lines indicate the origin of each dimension. The colours denote cluster membership, and 68% confidence ellipses (dashed lines) represent the spatial extent of each cluster.

Reaction Time data

Outliers defined as trials with average reaction times that were ± 3 SDs away (above or below) from the group mean ± 4 SDs away from the individual's mean performance were excluded prior to analysis. This represented a loss of 5.3% of the entire data set. For reaction times data, a linear mixed-effects model with condition, object pair, and their interaction as fixed effects, random intercepts and slopes for condition by participant was compared to a model without random slopes. Including random slopes significantly improved model fit, $\chi^2(5) = 28.95$, $p < .001$, so the model with random slopes was retained. A likelihood ratio test comparing models with and without the interaction term between motion condition and object pairings indicated that the interaction did not significantly improve model fit, $\chi^2(2) = 1.64$, $p = .44$. The simpler model

including only the main effects revealed no effect of motion condition, $\chi^2(2) = 2.83$, $p = .24$, and no effect of object pairings, $\chi^2(1) = 3.37$, $p = .07$. Thus, there was no evidence that reaction times were modulated by either object similarity or motion.

3.5.3 Discussion

Overall, the results of the similarity ratings and multidimensional scaling suggest that consistent motion influences the perceived similarity of the shapes of objects relative to objects presented with inconsistent motion. For the analysis of similarity ratings, the consistent movement patterns led to higher similarity ratings of object pairs than inconsistent movements, with static objects generally falling between the two. Importantly, these effects were shown for both same and different object pairings, suggesting that coherent motion promotes perceptual grouping across varying degrees of object similarity, and this effect of motion was observed in similarity judgements of object pairs with the same shape as well as pairs with different shapes. In the overall analysis, there was no overall benefit or cost in similarity ratings when objects moved consistently or differently relative to static objects.

The finding that objects with the consistent movement were rated as more similar than those with an inconsistent movement extended across objects pairs with the same shape as well as pairs with a different shape. However, for object pairs with the same shape, the ratings to their static versions did not differ from the ratings to either motion condition. In other words, the ratings to the static version of the same object shapes lay somewhere between those higher ratings to objects with the consistent movement and lower ratings to objects with inconsistent movements. Overall, the similarity ratings to objects with the same shape were relatively high, suggesting that differences in perceived similarity between static and moving versions may have been difficult to observe. Instead, we find evidence for a benefit of consistent movement on ratings over inconsistent movement which nevertheless suggests that motion, even when objects are the same shape, has an effect on object perception.

The result that objects motion affects perception even when shapes are different is also consistent with this conclusion. Although the perceived similarity decreased with greater step difference between object pairs, this effect was influenced by motion such that object pairs with inconsistent motion were rated as less similar than object pairs when the motion was consistent. As with the same object pairs, ratings to different object pairs that were not moving (static) did not differ from either of the movement conditions.

NMDS analyses further support this conclusion: the MDS configurations revealed more compact and separable clustering structures when object pairs moved in the same way, whereas clusters were fewer and more cohesive when motion differed or if the objects did not move (static). Finally, the absence of significant reaction time differences across conditions suggests that these perceptual effects are not due to decisional effort or task difficulty.

Our finding that motion cues are integrated into similarity judgement of object shapes are consistent with previous evidence (Vuong et al., 2012; Chuang et al. 2012; Jastorff et al., 2012) and add further support to the idea that object movement affects the representation of objects in memory (Newell, et al., 2004; Harrison et al., 2005; Wallraven & Bülthoff, 2009; Chung et al., 2023). Taken together, Experiment 2A provides evidence that consistent motion enhances perceived similarity and perceptual organization, over and above static appearance, but does not influence response latencies.

3.6 Experiment 2B

In Experiment 2B, we investigate whether cross-modal motion cues further modulate similarity judgments by introducing auditory motion cues. Specifically, we assessed whether the presence of sound correlated to visual object motion would lead to higher shape similarity judgments, given that crossmodal correlations can strengthen perceptual binding (Stein & Meredith, 1993; Lewis & Noppeney, 2010; Sanabria et al., 2004). We predicted: (i) replicating Experiment 2A, same-object pairs would be rated as more similar than different-object pairs, and

consistent motion would increase similarity relative to inconsistent or static motion (ii) similarity ratings would be further modulated by sound motion cues, such that the contrast between consistent and inconsistent visual motion would be larger in the presence of sound motion cues. (iii) Consistent visual motion, especially when combined with sound motion cues, would also lead to faster response times, reflecting efficient perceptual grouping. Alternative outcomes were also considered: sound motion cues might lead to higher similarity in a uniform way across motion conditions or exert no effect at all. As in Experiment 1A, motion effects are expected to be strongest for same-object pairings, with weaker or absent effects for different-object pairings.

3.6.1 Method

Participants

We recruited 61 adult participants (mean age = 42.4 years, SD 12.25; 34.4% female) using Prolific (<https://www.prolific.co/>), based on the following inclusion criteria: fluency in English, normal or corrected-to-normal vision and no hearing impairments. Most completed at least secondary level education (98.4%; with 69% completing third level education). All the participants were naïve to the purpose of the study and required to use a desktop computer to complete the experiment. They received payment for taking part at a rate of £9.00 per hour. The study was approved by the Trinity College Dublin, School of Psychology Research Ethics Committee and complied with EU data protection legislation (GDPR).

Design

The experiment was based on a within-subject, factorial design, with motion condition and sounds as the main factors. Similar to Experiment 2A, the motion condition had 3 levels, e.g., static, consistent or inconsistent movement. The sound condition had 2 levels: no sound or sound. Within each of the 3 object sets (see above), half of the object's pairs were randomly assigned either to the 'sound' condition and the other half to the 'no sound' condition across participants.

This meant that the maximum step size tested for each object was 5. We felt that this was a reasonable deviation from the design of Experiment 2A, since step sizes greater than 5 had little effect on similarity ratings (see Figure 18B). As in Experiment 2A, each object set was assigned to one of three motion conditions, and this assignment was counterbalanced across participants. Trials containing pairs of objects were blocked according to 3 motion conditions, with 84 trials per block (each object was compared once with itself and once with every other object in the subset, leading to $6 + (6 \times 5)/2 = 42$ trials, repeated twice for each participant yielding a total of 252 trials), and trial order was randomized within blocks.

Procedure

We adopted a similar procedure described in Experiment 2A, with two minor differences. First, informed by the results obtained from Experiment 2A, we used a continuous 7-point Likert scale instead of an ordinal type, in order to capture more nuanced results. Second, we introduced a sound check at the beginning of the block, in which asked participants were presented with three trials consisting of sound clips and were asked to report whether the tone they perceived was high or low in pitch. Failure to successfully complete (i.e., 2/3) these sound-check trials terminated the experiment because of a concern that sounds were not presented or heard. As per Experiment 2A, at the end of the experiment participants were asked to provide a brief description of how they had judged similarity among the objects. Each participant took approximately 30 minutes to complete the experiment. Experiment 2B was pre-registered prior to data analysis on the Open Science Framework, <https://osf.io/esz2b>, and further details, including the analysis script, are available at https://osf.io/78skg/?view_only=31bd486a1ef84bafaf96dd93981f0fd0.

3.6.2 Result

Participants accessed the experiment through different web browsers (86.8% via Chrome browser; 4.9% on Edge and 8.2% on Firefox). All participants completed the experiment as instructed, with full use of the scale. The data were analysed in the same manner as reported in

Experiment 2A. All participants provided a description, post-experiment of how they conducted the task: the words 'shape', 'shapes', 'looked' and 'tried' were the most frequently reported, whilst other spatial words (top, bottom, edges) were less frequently mentioned. As per from Experiment 2A, there was only one mention of object motion and sound in these reports, for further details refer to Supplementary Materials - Figure S7B.

Similarity rating data

Table 5

Mean normalised similarity ratings between pairs of objects across the sound and motion conditions for each step size in Experiment 2B.

Step size	Sound			No Sound		
	Static	Consistent	Inconsistent	Static	Consistent	Inconsistent
0	0.87	0.86	0.77	0.86	0.86	0.78
	<i>0.19</i>	<i>0.18</i>	<i>0.24</i>	<i>0.17</i>	<i>0.17</i>	<i>0.23</i>
1	0.71	0.82	0.70	0.70	0.82	0.68
	<i>0.28</i>	<i>0.20</i>	<i>0.27</i>	<i>0.27</i>	<i>0.19</i>	<i>0.26</i>
2	0.55	0.70	0.63	0.60	0.71	0.58
	<i>0.30</i>	<i>0.26</i>	<i>0.29</i>	<i>0.29</i>	<i>0.25</i>	<i>0.29</i>
3	0.45	0.63	0.57	0.53	0.65	0.52
	<i>0.30</i>	<i>0.29</i>	<i>0.31</i>	<i>0.31</i>	<i>0.26</i>	<i>0.31</i>
4	0.40	0.56	0.50	0.47	0.58	0.47
	<i>0.31</i>	<i>0.31</i>	<i>0.31</i>	<i>0.31</i>	<i>0.28</i>	<i>0.30</i>
5	0.33	0.52	0.43	0.40	0.50	0.43
	<i>0.28</i>	<i>0.33</i>	<i>0.31</i>	<i>0.32</i>	<i>0.29</i>	<i>0.31</i>

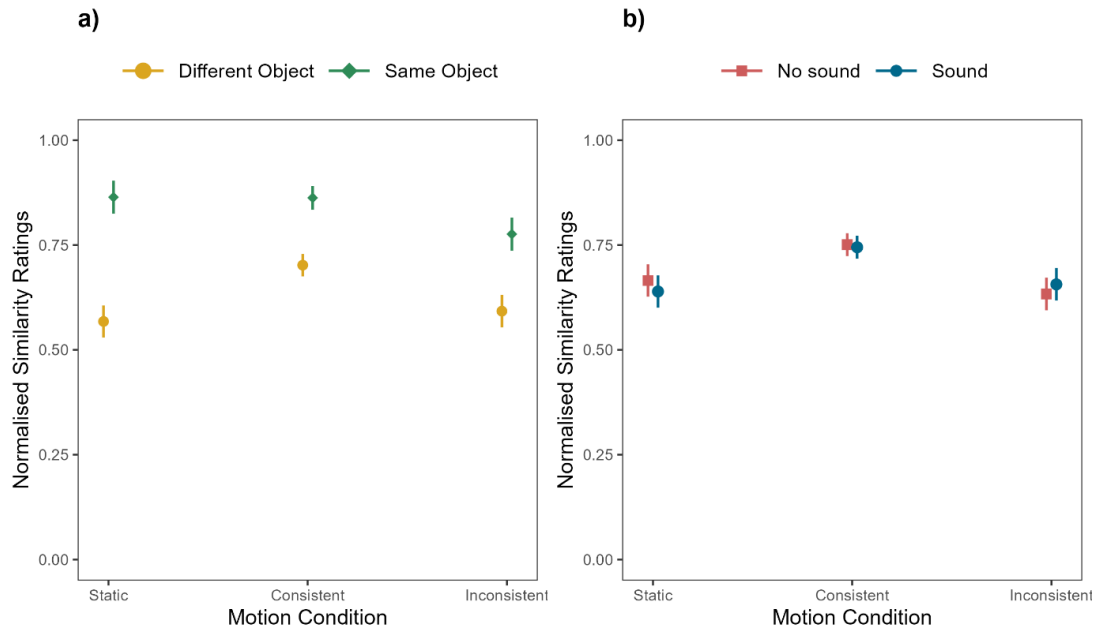
Note: Note that the original ratings are normalised (0-1) as described in the text. Standard deviations are shown in italics.

A Kruskal–Wallis test indicated a significant difference in similarity ratings across stimulus sets, $\chi^2(2) = 195.31, p < .001$. To account for this, stimulus set was included as a random intercept in all subsequent mixed-effects models. The mean similarity ratings for each of the sound, motion (static, consistent or inconsistent) and object (same, different) conditions are presented in Table 5. For similarity ratings data, we first checked for individual variability in the effect of both sound and movement across participants. The plot informed that the effect of movement condition (i.e.,

the slope) was not consistent across participants thus we decided to fit a linear mixed-effects regression model to the data in order to take this into account (and which is not possible with ANOVA). To check if a random slope would improve the model, we performed a comparison between models, with and without motion condition as random effect. A likelihood ratio test suggested that the motion condition as random effect did significantly contribute to the model predicting similarity ratings, $\chi^2(5) = 1539.7$, $p < .001$. Thus, we included motion condition as a random effect, in all the subsequent analysis. We found main effects of motion condition, $\chi^2(2) = 39.8$, $p < .001$, and object pairing, $\chi^2(1) = 2335$, $p < .001$, but not of sound condition, $\chi^2(1) = 0.68$, $p = .41$. We further investigated all 2-way interactions (sound condition by object pairing; motion condition by sound condition; motion condition by object pairing) individually, by comparing between models, with and without each 2-way interaction separately. A likelihood ratio test suggested that both the motion condition by object pairing, $\chi^2(2) = 196.5$, $p < .001$, and the motion condition by sound condition, $\chi^2(2) = 28$, $p < .001$ interactions significantly contributed to the model (as shown in Figure 20), while the sound condition by object pairing interaction did not contribute to the model predicting similarity ratings, $\chi^2(1) = 0.05$, $p = .82$. A comparison between models with and without the 3-way interaction (i.e. motion condition by object pairing by sound condition) was performed. A likelihood ratio test suggested that the 3-way interaction did significantly contribute to the model predicting similarity ratings, $\chi^2(2) = 13.71$, $p = .001$. The 3-way interaction suggested that in the 'sound' condition, the distinction in similarity ratings between different and same object pairings across static versus inconsistent motion was larger relative to the 'no sound' condition. However, the size of the contrast was small (EMM = .077, for detailed information refer to Supplemental Materials – Figure S9).

Figure 20

The plot shows the summary of the main findings from linear mixed-effects models predicting probability of similarity ratings in Experiment 2B.



Note: The plot in panel a) shows the interaction effect of motion condition by object pairing on similarity ratings. The plot in panel b) shows the interaction effect of motion condition by sound condition. Error bars represent 95% CIs.

In relation to object shape, ratings to different object pairs were consistently rated as less similar than same shape pairs, regardless of motion condition. Ratings to object pairs with the *consistent movement* were significantly higher than those to objects with *inconsistent movements* (EMM = 0.01, 95% CI [0.051, 0.146], $p < .0001$). Ratings to static objects were significantly lower than those to objects in with consistent motion (EMM = 0.07, 95% CI [-0.106, -0.0267], $p = .0004$). For the motion condition by object pairing interaction effect, post-hoc comparisons suggested that when object pairs were the same shape, the similarity ratings were significantly higher when presented in a static condition compared to when their movement was the inconsistent (EMM = 0.088, 95% CI [0.015, 0.161], $p = .007$), but did not differ from those in the consistent motion condition (EMM = 0.002, 95% CI [-0.052, 0.055], $p = 1$). Similarity ratings significantly differ

between the consistent and inconsistent motion conditions (EMM = 0.086, 95% CI [0.024, 0.149], $p = .0010$). When object pairs were of a different shape, static object pairs were rated as significantly less similar than object pairs moving in a consistent manner (EMM = -0.134 , 95% CI [-0.184 , -0.085], $p < .0001$), but ratings were the same across the static and inconsistent movement conditions (EMM = 0.025, 95% CI [-0.095 , 0.046], $p = 1$). Moreover, object pairs moving in a consistent way were rated as significantly more similar than ones moving in an inconsistent manner (EMM = 0.110, 95% CI [0.051, 0.169], $p < .0001$).

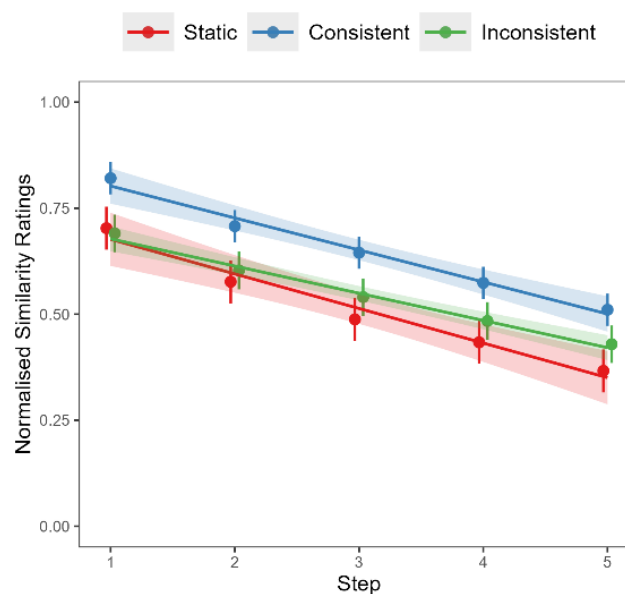
For the sound condition by motion condition interaction effect, a post-hoc comparison suggested that in the static condition objects presented without sounds (blue points in Figure 20b) were rated as significantly less similar than pairs without sound in the consistent movement (EMM = -0.056 , 95% CI [-0.107 , -0.005], $p = .019$). However, without sound the difference between static and inconsistent movement conditions did not reach significance (EMM = 0.056, 95% CI [-0.015 , 0.127], $p = .280$). The ratings to the objects pairs with the consistent movement were significantly higher than those with inconsistent movement (EMM = 0.113, 95% CI [0.052, 0.173], $p < .0001$). When sound was present, object pairs presented in the static condition were rated as significantly less similar than pairs with the consistent movement (EMM = -0.076 , 95% CI [-0.127 , -0.025], $p = .0003$) but not to pairs with inconsistent movements (EMM = 0.007, 95% CI [-0.064 , 0.078], $p = 1$). Finally, the ratings to the objects pairs with the consistent movement were significantly higher than those with inconsistent movement (EMM = 0.083, 95% CI [0.023, 0.144], $p = .0011$).

For pairs made of different object, we averaged similarity ratings across sound condition and analysed them. Due to violation of the homogeneity of variance assumption, linear mixed-effects models were fitted with random intercepts and slopes for condition per participant, instead of a standard repeated-measures ANOVA. A likelihood ratio test suggested that the motion condition by step interaction effect did significantly contribute to the model predicting similarity ratings,

$\chi^2(8) = 15.60$, $p = .049$, as shown in Figure 21. Post-hoc comparisons suggested that object pairs presented with consistent motion were rated as significantly more similar than the ones presented in the static condition at one step (EMM = -0.118 , SE = 0.029 , 95% CI [-0.222 , -0.014], $p = .0003$), two steps (EMM = -0.131 , SE = 0.029 , 95% CI [-0.235 , -0.027], $p < .001$), three steps (EMM = -0.157 , SE = 0.029 , 95% CI [-0.261 , -0.053], $p < .001$), four steps (EMM = -0.140 , SE = 0.029 , 95% CI [-0.244 , -0.036], $p < .001$) and five steps distance (EMM = -0.144 , SE = 0.029 , 95% CI [-0.248 , -0.040], $p < .001$). Moreover, object pairs presented with consistent motion was rated higher than the ones presented with inconsistent motion at one step (estimate marginal means = 0.130 , SE = 0.026 , 95% CI [0.037 , 0.224], $p < .001$), two steps (EMM = 0.104 , SE = 0.026 , 95% CI [0.011 , 0.198], $p = .0003$), three steps (EMM 0.105 , SE = 0.026 , 95% CI [0.011 , 0.198], $p = .0003$), four steps (EMM = 0.090 , SE = 0.026 , 95% CI [-0.004 , 0.183], $p = .0024$) and five steps distance (EMM = 0.081 , SE = 0.026 , 95% CI [-0.012 , 0.175], $p = .0069$). Differences between object pairs in the static and inconsistent motion conditions were not significant at any step (all $p > .15$).

Figure 21

The plot shows the main predictions from the linear mixed-effects models on the probability of similarity ratings for the motion condition across steps size in Experiment 2B.



Note: The solid line represents a weighted regression (lm) fit, and the shaded area indicates the 95% confidence interval around the smoothed trend.

Reaction Time data

Outliers defined as trials with average reaction times that were ± 3 SDs away (above or below) from the group mean ± 4 SDs away from the individual's mean performance were excluded prior to analysis. This represented a loss of 5.4 % of the entire data set. The mean RTs to each of the motion and sound conditions across object pairs are shown in Table 4.

Table 6

Mean response times(in milliseconds) to each of the sound and motion conditions across object step sizes Experiment 2B.

Step size	Sound			No Sound		
	Static	Consistent	Inconsistent	Static	Consistent	Inconsistent
0	1148	1105	1171	1157	1121	1121
	<i>546</i>	<i>539</i>	<i>540</i>	<i>544</i>	<i>549</i>	<i>549</i>
1	1175	1165	1197	1189	1098	1210
	<i>568</i>	<i>563</i>	<i>546</i>	<i>562</i>	<i>509</i>	<i>553</i>
2	1217	1139	1191	1199	1186	1219
	<i>581</i>	<i>581</i>	<i>570</i>	<i>583</i>	<i>574</i>	<i>597</i>
3	1225	1148	1203	1236	1122	1235
	<i>563</i>	<i>562</i>	<i>573</i>	<i>601</i>	<i>542</i>	<i>568</i>
4	1211	1106	1226	1231	1156	1192
	<i>573</i>	<i>521</i>	<i>587</i>	<i>599</i>	<i>591</i>	<i>542</i>
5	1202	1127	1209	1261	1140	1221
	<i>521</i>	<i>551</i>	<i>575</i>	<i>610</i>	<i>567</i>	<i>501</i>

Note: Standard deviations are shown in italics.

Prior to analysis of the reaction times data, we checked for individual variability in the effect of both sound and movement across participants. The plot informed that the effect of movement condition (i.e., the slope) was not consistent across participants thus we decided to fit a linear mixed-effects regression model to the data in order to take it into account, which was not possible with ANOVA. To check if a random slope would improve the model, we performed a comparison between models, with and without motion condition as random effect. A likelihood ratio test suggested that motion condition as random effect did significantly contribute to the model

predicting reaction times, $\chi^2(5) = 870, p < .001$. Thus, we included motion condition as error, in all the subsequent analysis. A comparison between models, with and without the 3-way interaction (motion condition by object pairing by sound condition) was performed. We found significant main effects of motion condition, $\chi^2(1) = 8.38, p = .015$, and of object pairing, $\chi^2(1) = 18.84, p < .001$. We found no effect of sound condition, $\chi^2(1) = 0.01, p = .900$, nor any of the following two-way interaction effects: motion condition by object pairing, $\chi^2(2) = 3.22, p = .200$; sound condition by object pairing ($\chi^2(1) = 1.89, p = .170$); motion condition by sound condition ($\chi^2(2) = 5.160, p = .076$). There was no evidence for a three-way motion condition by object pairing by sound condition interaction, $\chi^2(2) = 0.32, p = .854$. For the main effect of motion condition, a post hoc comparison suggested that participants were significantly faster in the consistent compared to the inconsistent movement condition (estimate marginal means = -71.7 ms, 95% CI [-133, -9.94], $p = .018$). However, RTs in the static condition did not significantly differ from response times to the consistent movement condition (estimate marginal means = 57.7 ms, 95% CI [-11.2, 126.7], $p = .13$), nor to the inconsistent movement conditions (estimate marginal means = -14 ms, 95% CI [-81.5, 53.61], $p = 1$). For the main effect of object pairing, we observed slower RTs for different shape compared to same shape object pairs (estimate marginal means = 31.7 ms, 95% CI [17.4, 46.0], $p < .0001$).

3.6.3 Discussion

In Experiment 2B we specifically examined whether sounds that were correlated with the movement of the object modulate shape similarity judgements. Consistent with Experiment 2A the results suggested that motion and object pairing influenced similarity ratings: participants judged objects as more similar when they moved in a consistent way than when the objects were static, and objects with the same shape were rated as more similar than those with different shapes. In contrast, we failed to find evidence that object sound affected similarity ratings overall or response times. The absence of a main effect of sound condition suggests that auditory

information, as presented in this task, was not sufficiently informative relative to visual information (shape or motion) about the objects. In particular, the lack of RT effects for sound is consistent with previous findings that auditory cues are often not integrated unless they are explicitly task-relevant or semantically informative (Bizley et al., 2016; Chen & Spence, 2010).

The results were qualified by two-way interactions both between motion condition and object pairing, and between motion condition and object sound. These interactions suggest that motion modulates the influence of both object identity and sound on perceived similarity. This is consistent with theories of object perception in which spatiotemporal congruency (e.g., shared motion) and structural features (e.g., shape) are dominant cues object recognition (Kaiser et al., 2019; Vuong & Tarr, 2004; Vuong et al., 2005). The process of object recognition inherently relies on assessing the similarity between a perceived object and stored representations in memory (Edelman & Bülthoff, 1992; Tarr & Pinker, 1989). The absence of a main effect of sound condition suggests that auditory information, as presented in this task, was neither sufficiently diagnostic nor strongly relevant into the visual comparison process. Although we observed a significant interaction between motion and sound, the effect size was small, and post-hoc comparisons revealed only subtle differences. This suggests that motion may slightly enhance the perceptual salience of sound, or vice versa, but this influence appears secondary to the effects of object shape and motion.

The interaction between motion and object pairing, seemed to be mainly driven by the ratings to the object pairs when their motion was the same: consistent with the results of Experiment 2A, ratings were relatively higher to the consistent motion than inconsistent motion conditions when the object pairs had both different and same shape. Moreover, ratings in the consistent motion were higher than static condition when the object pairs had different shape, and ratings in static condition were higher than inconsistent motion condition for object pairs made of same shape. This latter finding, of a relative increase of perceived similarity in static

condition (compared to inconsistent) for same shape objects and decrease for different shape objects, was not found in Experiment 2A, in which static condition was equally similar to both motion conditions across object pairs. This might be due to the modulatory influence of auditory information on perceptual similarity in Experiment 2B. When the object pair comprised different shapes, the addition of a congruent sound may have strengthened object-specific associations (i.e., the object moving in a consistent manner), leading to higher similarity ratings for the Consistent compared with the Static condition. By contrast, when object identity was already evident (same-object pairings), static information may have been treated as sufficiently reliable for perceptual judgements, in comparison with the Inconsistent condition. Nevertheless, with same object pairings, object motion increased similarity ratings relative to no motion. Results on reaction time data mirrored these patterns: we found faster responses for *Same Object* and *Consistent movement* trials than Different Object and Inconsistent movement trials, respectively, but no main effect of sound. These findings on reaction time effects of motion condition and object pairing were observed in Experiment 2B but not in 2A suggesting they may arise from differences in task demands or from the introduction of the sound dimension. However, as sound did not independently affect similarity judgments or RTs, cross-modal influences in this paradigm appear minimal.

Moreover, the analysis of similarity ratings across steps revealed a significant interaction between motion condition and step, indicating that the effect of motion is modulated by step size. Post-hoc contrasts showed that differences between Consistent and Static motion conditions were largest at smaller steps and reduced at larger steps, whereas differences between Consistent and Inconsistent motion decreased slightly across steps. Overall, these results suggest that consistent motion reliably increased similarity ratings, although the magnitude of this effect varied slightly with step. Importantly, the overall pattern remained the same across steps, with objects in the Consistent motion condition rated as more similar than those in both Static and Inconsistent

motion conditions, while objects in the Static condition were rated similarly to those in the Inconsistent motion condition.

Taken together, the results of Experiment 2B generally replicate the findings of Experiment 2A: object identity and motion congruency are the most reliable predictors of perceived similarity. Overall, the findings from Experiment 2B strengthen the conclusion that shape, and motion are dominant cues in object similarity judgements, with object sounds have a weak, modulatory role.

3.7 Experiment 3

In Experiments 2A and 2B we consistently found that object motion modulates shape-based similarity judgments between unfamiliar objects, even when participants were instructed to ignore motion, suggesting that there is a perceptual interaction between the dimensions of object shape and motion. This finding raises the question about whether shape and motion are independently encoded (separable features), or if they are integrated into a perceptual representation of the object (integrated features). To address this, in the following experiment we adapted the Garner interference paradigm (Garner, 1974) to conduct a speeded classification task. In this paradigm, interference is assumed when variation in a one irrelevant dimension (e.g., motion) disrupts the classification of information based on a relevant dimension (e.g., shape). If the two dimensions are separable, then variation in the irrelevant dimension should not affect performance. On the other hand, if these dimensions are integral, then variation even in the irrelevant dimension would lead to performance interference (i.e. increased response times and/or error rates). The Garner interference paradigm allows us to test this directly by comparing performance across three conditions. That is, by systematically manipulating the irrelevant dimension such that it is held constant (i.e. 'baseline'), or randomly varied (i.e. 'filtering'), or varied in a manner that is correlated with changes in the relevant dimension (i.e. 'correlation'), we can test whether perceptual processing of shape and motion is independent or integral.

3.7.1 Method

Participants

We recruited 35 adult participants (mean age = 38.79 years, SD = 11.29, 45.5% female) using Prolific (<https://www.prolific.co/>), based on the following inclusion criteria: fluency in English, normal or corrected-to-normal vision and no hearing impairments. The majority (87.9%) reported having completed second-level education, 76% held a third-level qualification or higher. Most accessed the study using the Chrome browser (66.8%), with the remainder using Edge (12.2%), Firefox (15%), Safari (3%), or other browsers (3%). All participants were naïve to the purpose of the study and were paid for participating (at a rate of £9.00 per hour). The study was approved by the Trinity College Dublin, School of Psychology Research Ethics Committee and complied with EU data protection legislation (GDPR).

Stimuli

The same object stimuli (static and dynamic versions) used in Experiment 2 were again used here. An additional object stimulus (a cube) was designed with identical properties and dimensions of the previous stimuli and was animated by applying the same motion patterns previously described (see Figure 16). All static and moving object stimuli are available on the Open Science Framework page for Experiment 3, https://osf.io/b5ez9/?view_only=c6bda5cb49c64e7d8a8b8ffb117dd309.

Design

The experiment was based on a 2x3 within subject design, with relevant dimension (shape or motion) and classification conditions (baseline, filtering, correlation) as factor. Both relevant dimensions are presented in a block design, which order is fully counterbalanced across participants. When the shape was the relevant dimension, the classification conditions were defined as follows: in the 'baseline' condition only the shape is varied while motion was held

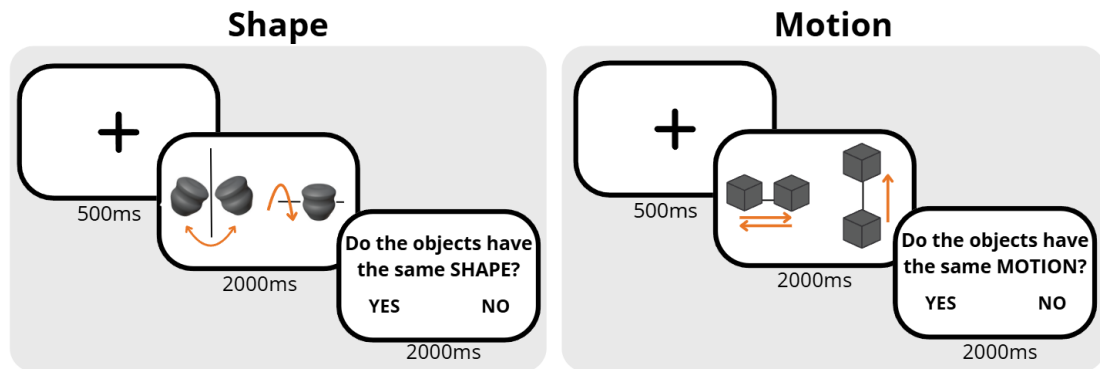
constant; in the 'filtering' condition both motion and shape were randomly varied; in the 'correlation' condition specific shapes were presented with specific motions. On the other hand, when motion was defined as the relevant dimension, the following occurred: in the 'baseline' condition only the motion pattern was varied while shape was held constant; in the 'filtering' condition both motion and shape were randomly varied; and in the 'correlation' condition specific shapes were presented with specific motions. Trials were blocked according to each of the relevant dimensions and block order was counterbalanced across participants. Each of the two blocks contained 384 trials (64 'baseline' trials; 264 'filtering' trials; 64 'correlation' trials) and the order of trials was randomized within the blocks, with a total of 768 trials. Participants took approximately 45 minutes to complete the whole experiment.

Procedure

All participants were first shown a study information sheet and asked to give their informed consent form to proceed, followed by a basic demographic questionnaire. The experiment started with five practice trials to allow participants to familiarise with the task. This practice was immediately followed by a test phase. In the test phase, participants were presented with a pair of objects in any one trial and were instructed, in advance, to focus on either the object's shape or the object's motion (how it moved). Their task was to judge, as accurately and as quickly as possible whether the object pairs had the same shape or (in a different block) motion. They provided their response by pressing 'a' on the keyboard to indicate 'yes, they have the same shape/motion' and 'l' to indicate 'no, they do not have the same shape/motion'. The response keys were counterbalanced across participants. A trial was structured as followed: a fixation cross appeared on the centre of the screen for 500ms, after which two stimuli (static or moving) were simultaneously presented, one on the left and one on the right side of the screen. The stimuli remained on the screen for a maximum of 2 seconds. Participants were then prompted for a response which was time limited to a maximum of 2 seconds when the trial was ended (see Figure 22).

Figure 22

An illustration of two trial examples from a speeded classification task where participants judge whether two objects share the same feature along a relevant dimension in Experiment 3.



3.7.2 Result

The data were analysed with R via RStudio (R Core Team, 2021). Post-hoc comparisons were conducted using the 'emmeans' package (Lenth Russell et al., 2022) and the Bonferroni correction was applied to correct for multiple comparisons. Experiment 3 was pre-registered prior to data analysis on the Open Science Framework, <https://osf.io/raksy>, and further details, including the analysis script, are also available at https://osf.io/b5ez9/?view_only=c6bda5cb49c64e7d8a8b8ffb117dd309. We first tested for Interference effects by comparing Baseline and Filtering conditions for both reaction times (RTs) and error rates. Interference effects were defined as the performance difference between the Filtering and Baseline condition: $GI = MRT_{\text{Filtering}} - MRT_{\text{Baseline}}$, with the same formula applied to error rates. Subsequently, to assess whether the covariation of shape and motion cues improved performance, we evaluated Redundancy Gains (RG). These were defined as the performance difference between the Baseline and Correlation conditions: $RG = MRT_{\text{Baseline}} - MRT_{\text{Correlation}}$, and likewise for error rates. Redundancy Gains reflect improved performance (faster RTs, higher accuracy) when both cues vary together.

Three participants were excluded from the final sample due to missing data points and technical issues during testing, leading to a final sample of thirty-three participants. Prior to statistical analysis, reaction time (RT) data were cleaned by excluding trials with latencies of RTs < 250 ms or > 2800 ms, which amounted to 0.05% of the total data set. The means of both RTs and error rates (in percentages) for each relevant dimension (motion and shape) across the experimental blocks (the baseline blocks were averaged) are reported in Figure 23. Trials were then aggregated by participant, condition, and relevant dimension to compute mean reaction times (MRTs) for correct responses and error rates (proportion of incorrect responses).

For reaction times, we fitted linear mixed-effects models (LMMs) using RTs from correct trials as the dependent variable. All models (including for error rates) included classification Condition, Relevant Dimension, and their interaction as fixed effects, and a random intercept for Participant. Random slopes for classification Condition were included where convergence allowed. To test for the presence of an Interference effect on reaction times, a comparison between models, with and without Relevant Dimension by classification Condition was performed. A likelihood ratio test showed that including the interaction significantly improved model fit, $\chi^2(2) = 34.35$, $p < .001$. The full model revealed significant main effects of both Relevant Dimension, $\chi^2(1) = 1286$, $p < .001$, and Condition, $\chi^2(2) = 6.84$, $p = .033$, as well as a significant interaction (presented in Figure 23A). For the main effect of Relevant Dimension, a post hoc comparison suggested that participants responded significantly slower overall when the relevant dimension was Motion compared to Shape. When Motion was the relevant dimension, responses were significantly slower in the Baseline compared to Correlation (EMM = 70.34 ms, 95% CI [26.85, 113.83], $p < .001$), and not significant different from Filtering (EMM = 33.11 ms, 95% CI [-7.44, 73.66], $p = .248$). Additionally, response times in the Correlation compared to Filtering condition (EMM = -37.23 ms, 95% CI [-77.97, 3.51], $p = .110$) were not statistically different.

For accuracy, we fit generalized logistic mixed-effects models (GLMMs) with a binomial link function (correct = 1, incorrect = 0) with and without the interaction term between Classification Condition and Relevant Dimension. A likelihood ratio test indicated that the interaction significantly improved model fit, $\chi^2(2) = 8.97$, $p = .011$. The model showed a significant main effect of Classification Condition, $\chi^2(2) = 7.72$, $p = .021$, but no significant effect of Relevant Dimension, $\chi^2(1) = 1.07$, $p = .301$. Despite the overall model significance, post hoc comparisons revealed no significant pairwise effects after Bonferroni correction, with all contrasts having adjusted p-values $\geq .07$. For instance, the largest difference (i.e. between Motion Correlation and Shape Filtering) had an odds ratio of 2.30 (95% CI [1.19, 4.45]), but this did not survive correction ($p = .013$ uncorrected, $p = .127$ adjusted).

Finally, we evaluated whether participants traded speed for accuracy, we computed Pearson correlations between mean RTs and error rates across participants, as reported in Table 5. A Pearson correlation was conducted to examine the relationship between mean reaction time (mRT) and error rate across participants. The analysis revealed no significant correlation between mRT and error rate: $r = -0.27$, $t(31) = -1.59$, $p = .122$, 95% CI [-0.56, 0.08], suggesting no evidence of a speed–accuracy trade-off.

Table 7

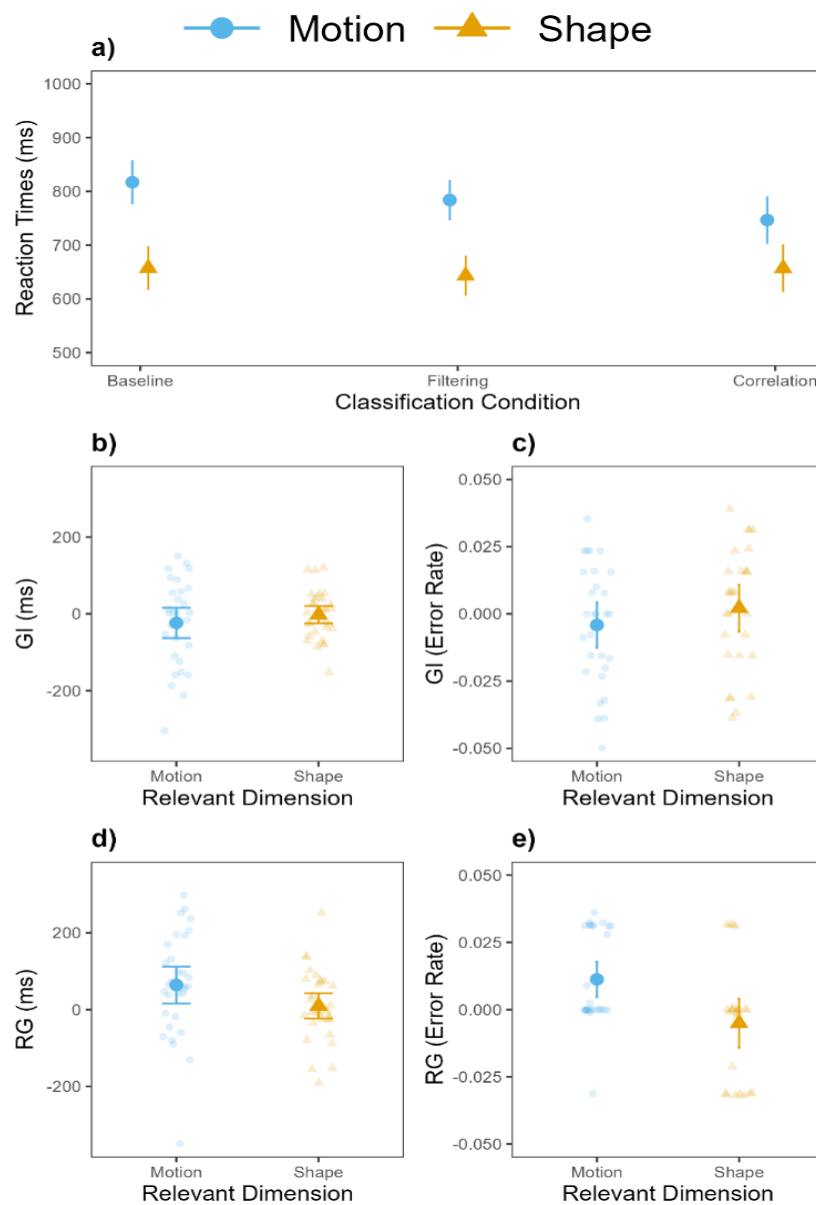
Mean error rates for each classification condition (baseline, correlation, filtering) and relevant dimension (shape, motion) as percentages in Experiments 3.

Relevant dimension	Classification condition		
	Baseline	Correlation	Filtering
Shape	2.29% (15.0)	3.20% (17.6)	3.81% (19.1)
Motion	3.77% (19.1)	1.72% (13.0)	3.34% (18.0)

Figure 23

Summary plots performance and interference effect of reaction times in milliseconds

in Experiment 3.



Note: Panel a) Main findings from linear mixed-effects models, showing the interaction effect of Relevant Dimension and Classification Condition. Error bars represent 95% CIs. Garner Interference (GI) Effects Across Relevant Dimensions (Shape vs. Motion) shown for b) the mean reaction times (mRTs) in milliseconds across and c) mean error rates. Redundancy Gain (RG; i.e., Baseline - Correlation) shown for d) the mean reaction times (mRTs) and e) error rates. All plots

distinguish between classification conditions where either Shape (orange) or Motion (blue) was the relevant dimension

3.7.3 Discussion

In Experiment 3, we used an adapted Garner interference paradigm to examine whether shape and motion are processed as separable or integral dimensions in the perception of novel objects. Across all classification conditions, participants were consistently slower when object motion was the relevant dimension compared to object shape. The presence or absence of interference effects varied systematically by dimension. RTs were faster in the Correlation than in the Baseline condition suggesting that a congruent (shape–motion) pairing facilitated processing relative to when only Motion varied (shape was the same). Together, these findings suggest that motion is an integral dimension to an object’s shape. In contrast, when shape was the relevant dimension, both interference and redundancy gain effects were minimal or non-significant, suggesting that the perception of object shape was not affected by motion variation, consistent with separable dimensional processing. Accuracy rates were high across all conditions, with only subtle differences across dimensions or conditions.

Garner’s (1974) framework posits that interference from irrelevant dimensions occurs when the perceptual system processes dimensions integrally, as fused or co-dependent, while separable dimensions can be selectively attended to without cost (Garner, 1974; Garner & Felfoldy, 1970). Our findings clearly demonstrate asymmetric separability: while motion is integrable with shape, such as performance suffers when irrelevant shape varies and improves when shape covaries with motion; shape appear to be separable from motion, leading performance to be largely unaffected by motion variation, whether correlated or not. This asymmetry is consistent with prior work on dimensional interactions in perception, where one dimension is more dominant or perceptually ‘anchored,’ while the other (often dynamic) is processed in a more context-sensitive or interactive manner (Garner & Felfoldy, 1970; Ashby & Maddox, 1994).

3.8 General Discussion

Across three experiments, we investigated how motion, shape, and sound contribute to perceptual similarity of novel objects. Through shape similarity judgements (Experiments 2A and 2B) and a speeded classification task (Experiment 3), we found that motion systematically modulates similarity structure and classification performance, whereas shape remains a more stable and separable dimension. Our results suggest that motion is not only perceptually salient but also integrally represented in object perception, particularly when shape information is ambiguous or impoverished.

In Experiment 2A, objects sharing consistent motion led to significantly increased perceived similarity between both same and different objects pairings compared to objects with inconsistent movements, even though participants were instructed to base judgments on shape alone. Our finding that motion cues are integrated into similarity judgement of object shapes are consistent with previous evidence (Vuong & Tarr, 2006; Vuong et al., 2012; Chuang et al. 2012), provide further support to the idea that object movement being integrated into perceptual representations in memory (Newell, et al., 2004; Harrison et al., 2005; Wallraven & Bülthoff, 2009; Chung et al., 2023), and expand findings showing that motion enhances structure recovery and perceptual grouping (Johansson, 1973; Ullman, 1979; Chuang et al., 2012). The observed clustering patterns from NMDS further revealed that shared motion induced more coherent perceptual groupings. This suggests that motion, like shape, can serve as feature in the representational space that supports similarity-based categorisation. Experiment 2B further extended this by introducing auditory cues correlated to motion, investigating whether multisensory cues modulate shape similarity judgements. Consistent with Experiment 2A, consistent motion led to higher similarity ratings, and while sound had no main effect on judgment, its interaction with motion suggests a modulatory role, consistent with the

multisensory literature highlighting temporal synchrony and semantic relevance as necessary for cross-modal integration (Spence & Driver, 2004; Giard & Peronnet, 1999).

Moreover, motion modulates the influence of both object identity and sound on perceived similarity, which aligns with evidence that spatiotemporal congruency and structural features are dominant cues object recognition (Vuong & Tarr, 2004; Vuong et al., 2005; Coste et al., 2021). The lack of main effect of sound might relate to its not informativeness relative to visual cues of objects, consistent with previous findings suggesting that auditory information is integrated if they are closely aligned with task-relevant features or semantically informative (Bizley et al., 2016; Chen & Spence, 2010). Together, findings from Experiment 2A and 2B support the view of motion influencing the structure of perceptual similarity spaces and affecting how objects are grouped and compared, even under conditions where motion is not the task-relevant. This integration likely reflects the use of motion as a diagnostic cue for structural inference, grouping, and identification (Johansson, 1973; Cutting & Kozlowski, 1977; Vuong et al., 2012).

Our use of highly homogeneous, novel 3D stimuli likely engaged subordinate-level processing, increasing the salience of fine-grained structural features and making recognition more viewpoint-dependent (Jolicoeur et al., 1986; Edelman, 1998; Foster & Gilson, 2002; Newell, 1998; Newell et al., 2001). In this context, motion can help to resolve structural ambiguity and facilitate identification, even when motion is incidental and not learned. These results align with prior work on ‘spatiotemporal signatures’ of dynamic objects, where observers encoded not only static shapes but also learned motion trajectories as integral to object identity (Stone, 1998, 1999; Liu & Cooper, 2003). Although our studies did not involve learned or repeated motion exposure, our findings suggest that even incidental motion patterns contribute to similarity perception.

In Experiment 3, we assessed dimensional separability of shape and motion dimensions in the perception of novel objects. Results showed asymmetrical interference: when motion was relevant, both interference (Filtering > Baseline) and facilitation (Correlation < Baseline) were

observed, consistent with integral dimensional processing (Garner, 1974; Ashby & Maddox, 1994). When shape was relevant, performance remained stable, consistent with separable processing, suggesting that shape is processed independently of motion, consistent with separable dimensional processing. These findings support the idea that dimensional separability is asymmetric, a pattern observed in other domains such as face and gender perception (Garner & Felfoldy, 1970; Ganel & Goshen-Gottstein, 2004). This asymmetry might reflect differences in processing demands, while shape is spatially stable and rapidly encoded; motion is transient, requires attentional tracking and might need cognitive resources, which aligns with previous literature showing that dynamic cues often demand sustained attention (Blake & Shiffrar, 2007; Scholl, 2009). This pattern aligns with findings showing that learned motion—particularly non-rigid motion—enhances recognition when shape information is degraded or ambiguous, or when viewpoint changes reduce structural reliability (Stone, 1998; Chuang et al., 2012; Vuong et al., 2009).

Moreover, these findings provide evidence for dimensionality modulation in object perception, extending the theoretical framework of perceptual restructuring beyond category learning tasks (Folstein et al., 2013). The integration of motion cues into shape similarity judgments, even when participants were instructed to focus on shape alone (Experiments 2A and 2B), shows that the brain reconfigures dimensional relevance based on available perceptual information (Wallraven & Bülthoff, 2009; Newell et al., 2004) to optimize perceptual performance (Kalish et al., 2005; Sewell & Lewandowsky, 2011). Particularly, the asymmetrical interference pattern observed in Experiment 3 might mirror the selective stretching of category-relevant dimensions reported in previous studies (Goldstone, 1994; Gureckis & Goldstone, 2008; Garner & Felfoldy, 1970). The compact clustering patterns arising from consistent motion suggest that dynamic information serves as a diagnostic cue for perceptual grouping, effectively creating a new organisational principle in the representational space (Johansson, 1973; Ullman, 1979). This dimensional modulation appears stronger when shape information is ambiguous or impoverished,

consistent with the adaptive nature of perceptual restructuring that prioritizes the most informative features for object discrimination (Goldstone et al., 1996; Notman et al., 2005).

The effect of motion may be modulated by the homogeneity of the stimulus set. As reported in prior work, homogeneous stimulus sets tend to promote subordinate-level categorisation, increasing reliance on fine-grained shape details and increasing viewpoint-dependence (Edelman, 1994; Tjan & Legge, 1998; Jolicoeur et al., 1986). In this context, dynamic cues like motion may become more salient, being used to disambiguate highly similar exemplars or stabilise identity across changes in view (Foster & Gilson, 2002; Chuang et al., 2012). Our use of novel objects might have activated such subordinate-level processing, increasing the influence of motion on perceived similarity. Our results show that dynamic (motion) and to a lesser extent auditory cues, change the perceptual space in which object representations and category structures emerge, potentially increasing both the flexibility and robustness of category learning mechanisms.

One limitation in this study might be that motion was not explicitly learned or task-relevant, and although prior research showed stronger effects for learned motion sequences (Stone, 1998; Chuang et al., 2012; Newell, 2010), our results suggest that even untrained motion can shape perceptual similarity and consequently categorisation. Future research could explore how incidental motion becomes encoded and how attentional focus, learning, or task relevance modulates this process. Moreover, while auditory cues had minimal impact here, other crossmodal pairings (e.g., touch, texture) may yield stronger integration.

3.9 Conclusions

In summary, our findings show that object motion is integrated into the representation of object shape in memory, with clear evidence for integral processing of motion and separable processing of shape, while auditory cues played a minor role under these task conditions. Taken together, the three experiments support the view of how shape and motion contribute to

perceptual similarity, and in turn to categorisation. Our findings have important implications on our understanding of how inter-object similarity is determined from cue combination and the consequent role of shape and motion features in the perception and categorisation of novel objects.

Chapter 4

**Visual and tactile motion cues enhance the
categorisation of novel object shapes: evidence from
remote testing using mobile devices**

4 Visual and Tactile Motion Cues Enhance the Categorisation of Novel Object Shapes: Evidence from Remote Testing Using Mobile Devices

4.1 Abstract

Categorisation is a fundamental cognitive process, involving the integration of information across the senses. We investigated remotely using smartphones whether visual and tactile motion cues could enhance object category learning and generalisation to novel object shapes. Two categories of similar shapes were associated with specific correlated visual and tactile vibration motion cues. After learning object categories, participants were assessed on categorisation of learned and novel objects across four cue conditions: shape-only, shape-visual motion, shape-tactile motion, and shape-visual and tactile motion. We also assessed if accuracy was influenced by blocked versus interleaved cue-conditions at test. In Experiment 4, we found more accurate categorisation and generalisation when all cues were available at test. In Experiment 5 we replicated this effect even when the reliability of the shape-only cue for predicting category membership was reduced. In Experiment 6, we found that the absence of motion cues during learning removed the benefit of motion cues at test. Overall, our findings suggest that multisensory motion cues benefit the formation of novel object categories and allow for better generalisation. The results have implications for our understanding of the underlying dynamic and multisensory nature of object categories and the predictive role of multisensory features on category formation.

4.2 Introduction

Imagine enjoying a peaceful wander through a forest when suddenly, to your horror, something lands on your back. After a while, you recognise that it is an insect and not a leaf falling from a tree. When trying to categorise the ‘something’ as an insect and not a leaf, you might rely on multiple cues, such as the feel of the flutter of the insect’s wings or the scuttling of its legs, while it moves into view. Such a task leverages multisensory integration for object categorisation learned through experience. Studies have shown that object category learning is a complex process involving the integration of multiple features to differentiate between object categories as well as generalise from individual exemplars (Goldstone & Hendrickson, 2010; Pérez-Gay et al., 2017). Shared object features, such as shape, sound, or movement, are used for category formation and for generalisation to novel objects. However, even though much is known about unisensory visual (e.g. Rosch, 1978) auditory (Griffiths & Warren, 2004; Feng et al, 2021; Brunel et al. 2013) and tactile (Newell, 2004) features on object categorisation, less is known regarding how multisensory cues are used together for category learning and generalisation (Newell et al., 2023).

Previous studies have shown that both vision (see Peissig & Tarr, 2007; Grill-Spector, 2003) and touch (Gibson, 1962; Lederman and Klatzky, 1987; Lederman & Klatzky, 1990) are sufficient to enable the learning and subsequent recognition of individual objects (Newell et al., 2001; Lacey et al., 2009). Moreover, object information can be efficiently transferred between vision and touch (Yildirim & Jacobs, 2013). For familiar objects in particular, the resulting shared representations may lead to the formation of more robust object categories than unisensory information alone (Yildirim & Jacobs, 2013; Gaissert & Wallraven, 2012; Haag, 2011; O’Callaghan et al., 2018; Broadbent et al., 2020), highlighting the crucial role of multisensory integration in the representation of object categories in memory (Naci et al., 2012). Despite our knowledge that multisensory features contribute to categorisation, it is unclear how multiple sensory inputs affect

the formation of novel object categories, and whether such multisensory categories allow for generalisation to novel exemplars. Indeed, previous studies investigating generalisation from learned multisensory representations have produced mixed results, with some reporting evidence for a benefit of multisensory learning on generalization (Wu et al., 2021), while others report no specific benefit of multisensory information on categorisation performance (Edmunds et al., 2020; Roark et al., 2021; Sun et al., 2023; Atkin et al., 2023; Roark et al., 2023; Li & Deng, 2023; Roark, 2024; O'Dowd et al., 2025). Some of the discrepancies across these studies may be due to the amount of prior knowledge of the object stimuli, the extent to which sensory information is correlated across modalities during learning, or the relative distinctiveness or predictability of each sensory cue to category membership.

Objects in the real world are not only multisensory, but often dynamic. Furthermore, when information from two sensory modalities shares temporal properties, such as synchronous movement, they are more likely to be combined (Parise & Ernst, 2016), which may serve to enhance any benefit of multisensory information for category learning. Within such a multisensory process, information gathered about the movement of an object through vision (Robert et al. 2023; Shatek et al. 2022) or touch (Gaisser & Wallraven, 2012, Simões-Franklin et al., 2011; Sumser et al., 2024) can play a fundamental role in category formation. Visual object movement can facilitate the recognition of novel objects (Stone, 1999; Newell et al., 2004; Setti & Newell, 2010) compared to static conditions, and the neural substrates underpinning tactile and visual object motion appear to be shared (Chan et al, 2010; Amemiya et al., 2017). Within the tactile modality, dynamic motion cues are often received in the form of vibrations, which can provide valuable information about specific object properties. In this regard, studies on vibrotactile discrimination suggest that touch can enhance discrimination (Mahns et al., 2006; Verrillo et al., 1969), particularly when visual temporal cues alone are unreliable or ambiguous (e.g., Pomper et al., 2014; Hirst et al., 2025). Tactile stimulation is quickly detected on the skin, and passive exposure to moving tactile information, such as vibrations or flutterings, can help

discriminate objects (Fleming et al, 2013; Fleming, 2017; Ryan et al., 2021; Shao et al., 2016; Ziat, 2023). Because visual and tactile motion cues contribute to object recognition and discrimination, we can assume they also contribute to object categorisation, but this has not hitherto been investigated. Building on this, the present work investigates whether motion-related cues can facilitate not only category learning but also generalisation across novel exemplars.

4.3 Outline of the Experiments

Across three experiments, we aimed to investigate whether tactile and visual object motion influenced the categorisation of novel object shapes, and if the formation of these multisensory categories subsequently benefited generalisation to new exemplars. In the first two experiments, participants learned to categorise novel object shapes presented with visual and tactile motion cues (in Experiment 6 only the object shape was presented during learning). We manipulated the cue informativeness of category membership, such that in Experiment 4 all cues were fully informative, while in Experiment 5 one cue (e.g., shape similarity) was less informative while both visual and motion cues remained fully informative. Visual motion consisted of each object either rolling, jumping, swinging, or shaking, whereas tactile motion consisted of vibrations that were correlated with the temporal characteristics of each of these visual motion patterns. For example, if an object was seen swinging from left to right, the tactile vibrations would share the same onset and offset events present during the visual motion sequence. Following successful learning, participants were then tested on their categorisation accuracy to learned exemplars and generalisation to novel exemplars. The test phase assessed performance across four different cue conditions in which we manipulated the number and type of cues available. Importantly, in our experiments, fewer motion than shape features were used as predictive cues to category membership. That is, each category was defined by nine shapes and two distinct motion features (e.g., rolling, swinging). As such, the within-category variability of the shape cue was relatively higher than within category motion cues. This relative variability across shape and motion cues

may result in different weightings of shape or motion cues for categorisation and generalisation performance, which we investigate here.

We conducted two different versions of our experiments: following learning, participants were tested on their categorisation of learned and novel objects in trials that were either blocked by cue condition or interleaved. A secondary aim of the current study was to understand the impact of blocked versus interleaved trial presentation on multisensory category learning. Previous studies have indicated that blocked versus interleaved presentation influences categorisation performance (Kornell & Bjork, 2008; Abel et al., 2021). Blocked trial presentation facilitates learning by reinforcing within-category featural similarities, whereas interleaved trial presentation promotes discriminative contrast across categories (Carvalho & Goldstone, 2015; Kost et al., 2015). Furthermore, it has also been suggested that interleaving trials improves long-term retention and category induction across sensory modalities (Ge et al., 2021; Abel, 2023). Thus, in multisensory contexts, interleaving of trials may further strengthen cross-modal cue integration, enhancing participants' ability to differentiate between categories (Abel, 2023). On the other hand, interleaving trials may disrupt performance if the ability to adapt to varying cue types across trials is not sufficiently flexible (Carvalho & Goldstone, 2019) and generate uncertainty about which cue to prioritise. Moreover, switching between different sensory modalities may increase cognitive load. Accordingly, the presentation of blocked or interleaved trials allowed us to tease apart these cognitive effects on cue integration.

4.4 Experiment 4

Experiment 4 was designed to explore whether the concurrent availability of visual and tactile motion cues enhances object category formation and categorisation performance. Building on prior research on the perception of moving objects (e.g. Setti & Newell, 2004; Chan et al., 2010), we hypothesised that the integration of visual and tactile motion cues with object shape would improve object categorisation and generalisation to novel exemplars relative to static shape

alone. We therefore predicted higher categorisation accuracy, including for novel exemplars, in conditions in which all cues were available relative to conditions in which one or more sensory cues were missing. To allow us to explore our secondary factor (blocked versus interleaved trials), we tested this prediction using trials that were either blocked by cue condition (Experiment 4A) or interleaved (Experiment 4B).

4.4.1 Method

Participants

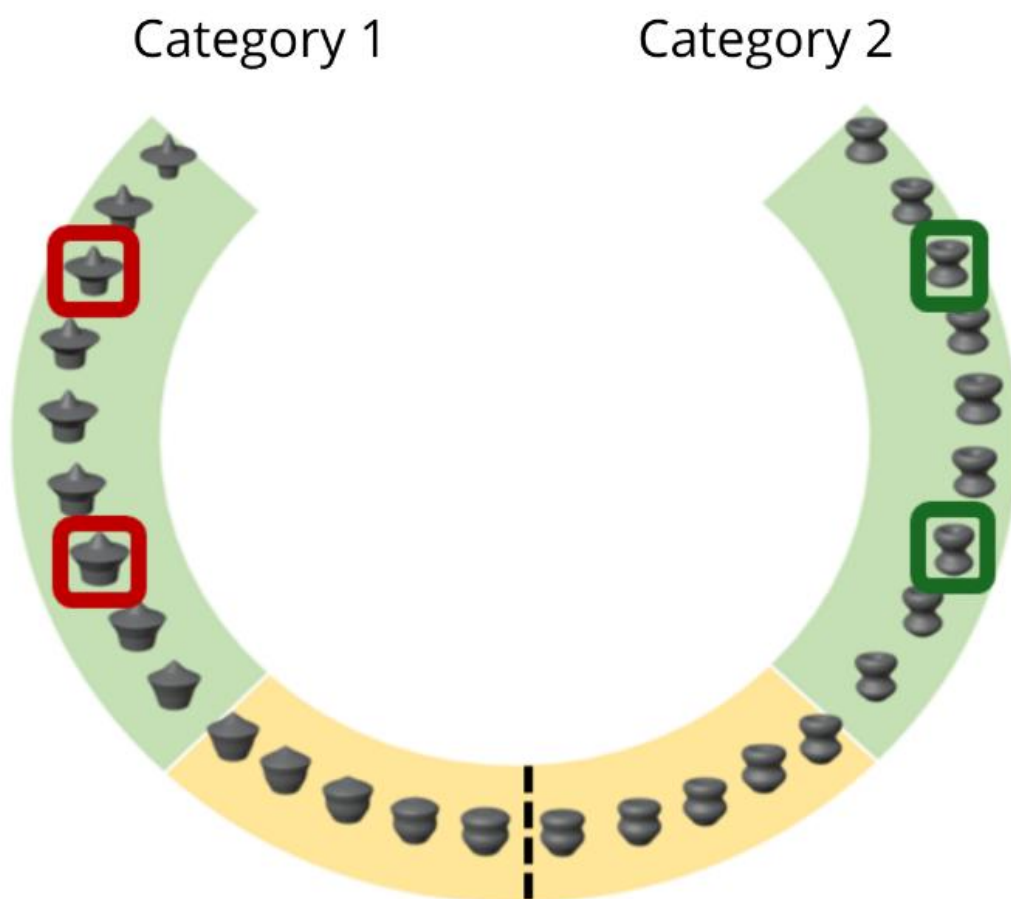
All participants were recruited using Prolific (<https://www.prolific.co/>), based on the following inclusion criteria: fluency in English, normal or corrected-to-normal vision and no hearing impairments. An a priori power analysis for each of the groups was performed in PANGAEA (Westfall, 2016; v0.2) for a within-subjects design. The advised minimum sample size is for 80% power to detect the effects of interest with size (Cohen's d) of 0.3, was 32 participants per experimental design (blocked or interleaved). We initially recruited 155 participants of whom 110 (71%) successfully reached the learning criteria. Of these participants who learned the categories, a further 33 failed to perform above chance in the test (representing 29% of learners) and their data were not included in any subsequent analyses. While this attrition rate may appear high, it is in line with findings from other studies on successful categorisation performance in laboratory settings (e.g. Smith et al., 2014; Roark & Chandrasekaran, 2023). In total, 77 participants (mean age = 38.83 years, SD = 10.97; 52 % female) completed the experiment online with 45 participants allocated to the blocked design (Experiment 4A) and 32 to the interleaved design (Experiment 4B). All participants were naïve to the purpose of the study and were compensated at a rate of £9.00 per hour. The study was approved by the Trinity College Dublin, School of Psychology Research Ethics Committee (approval number SPREC102020-50) and complied with GDPR data protection legislation.

Stimuli

We created 28 novel object shapes (3D-designed) with uniform spacing between neighbouring objects (except for object shapes at each end of the shape space, as shown in Figure 24), see Chapter 3 (General Stimuli section) for a detailed description of the 3D modelling procedure followed. These object shapes were used as stimuli throughout all the experiments. All object stimuli and scripts for generating the visual masks are available on the Open Science Framework [page](https://osf.io/s369c/?view_only=9015e8935be24628b5b15596b8eb6271) for Experiment 4, https://osf.io/s369c/?view_only=9015e8935be24628b5b15596b8eb6271.

Figure 24

An illustration of the 'shape space' of object stimuli used in the Experiments 4, 5 and 6.



Note: The shape space was adapted from a previous circular shape space described by Li et al., (2020). The objects are arranged within the shape space such that neighbouring objects are equal in angular distance from each other. The exception to this is the object stimuli at the two

endpoints of the shape space. Participants were trained to categorise objects highlighted in green and tested using both these learned as well as a set of novel object stimuli highlighted in yellow. The dashed line depicts the (arbitrary) category boundary of this shape space used in the Experiments. The learned stimuli are highlighted in green (9 per category) whereas the stimuli used to test for generalisation are highlighted in yellow (5 per category). The green and red framed stimuli are relevant to the design of Experiment 2 only. See Methods for further details.

We allocated an arbitrary category boundary to the middle of the stimulus shape space (see Figure 24), such that 14 objects were allocated per category. Within each category, shapes were highly similar, and shape similarity was the main cue for categorisation. Of the 14 objects per category, 9 objects were selected at the extreme point of the category for the learning session (green highlighted object shapes in Figure 24) and 5 objects per category, closer to the category boundary, were used for generalization testing (yellow highlighted object shapes in Figure 24). The use of object stimuli nearest to the category boundary for generalisation was to avoid using clearly distinguishable shapes, which would have reduced the difficulty of the task and possibly obscured any generalisation effects.

Each rendered 3D object was animated based on one of four different visual motion patterns; swing, jump, roll or shake, using the Workbench Engine. The motion patterns were chosen based on type (e.g., smooth or abrupt movement) and reference (e.g., movement of the object's horizontal or vertical axis). All movement sequences had a duration of 2 seconds. For the animation, each object stimulus was presented against a background consisting of two grey walls and a grey floor (using Blender). The camera angle (coordinates: 10.9 m, -14m, 4m; rotation: 70°, 0°) and lighting (point, radius 0.1m, 1000W; coordinates: 11m, -14 m, 6.9m; rotation: 40°, 34.8°) were both held constant. The light source illuminated the scene from above, creating an object shadow on the floor to aid depth and motion perception. The individual moving objects were then extracted and used as stimuli in the experiment to be displayed against a black background on a

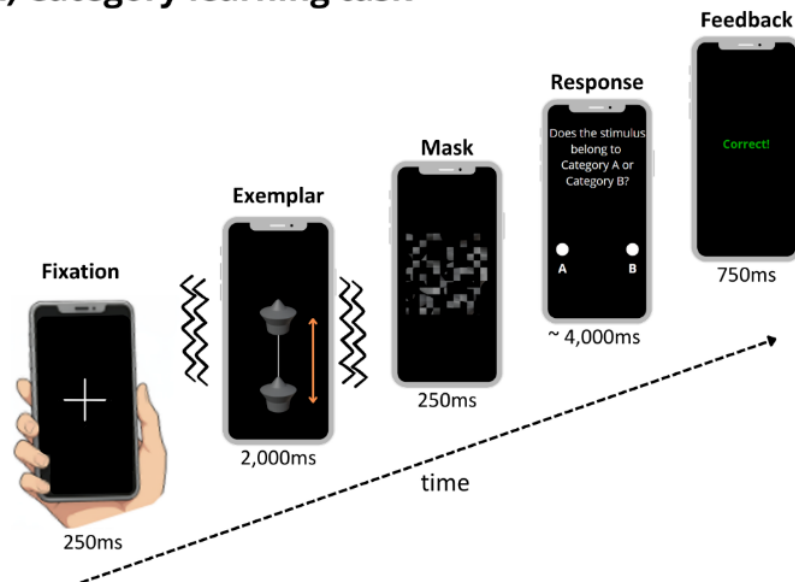
mobile phone screen. The entire screen was used to display the objects and response options (see Figure 25 for an illustration).

Tactile stimuli were delivered via the `navigator.vibrate()` function; a method of delivering tactile stimuli remotely via android smartphone browsers. This method is described, tested and validated elsewhere (Hirst et al., 2025), in short it allows the presentation of a vibration pulse sequence by specifying binary on/off commands, it does not allow for manipulation of vibration amplitude. Our tactile stimuli therefore consisted of a vibration pulse sequence derived from the visual motion pattern of each stimulus. This was achieved by extracting the temporal dynamics of visual motion for each movement (i.e. frequency and amplitude of movement) and mapping this onto an audio waveform with corresponding frequency and amplitude. The audio waveform was analysed using a 100ms rolling average to smooth fluctuations in amplitude, and a threshold was applied at 40% of the maximum amplitude to detect markers used to define pulse onset (e.g., 'on' or vibration) and offset (e.g., 'off' or pause) events. The resulting vibration patterns therefore closely mirrored the structure and timing of the visual motion stimuli with high temporal precision. A custom Python code (available on the OSF project page) ensured that the output format retained the timing and structure of the original input and was compatible with PsychoPy's tactile delivery system. The validity of our tactile motion stimuli was supported by the results of pilot tests requiring participants to match vibration pulse sequences with their visual counterparts as well as an assessment of synchrony judgements based on tactile-only and bimodal (i.e., visual motion paired with tactile vibration) stimuli (see Supplementary Materials – Table S7 for details of the results).

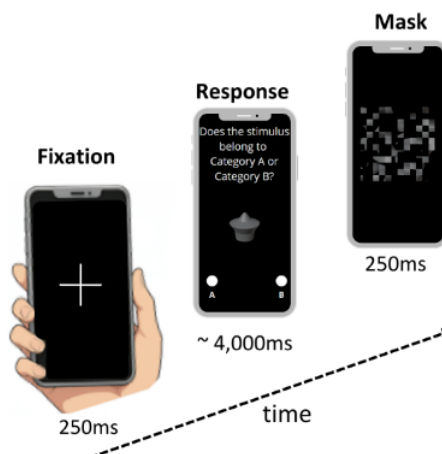
Figure 25

A schematic illustration of the sequence of events from stimulus presentation to the response display used in (A) the learning session and (B) the categorisation test in Experiments 4 and 5.

A) Category learning task



B) Categorisation test



Note: Only android mobile phones were used to in the experiment. The timeline of events starts from the left image and proceeds towards the right image for each of the learning and test examples. Wavy lines represent tactile vibrations which were correlated with the visual motion patterns. The categorisation test (B) depicts a trial from the Shape-only (Sv) condition. See text for details on the trial structure.

Design

The experiment was structured around two main sessions: a learning block (with feedback) followed by a test block (without feedback). We conducted two versions of the experiment in which trials were either blocked (i.e., Experiment 4A) or interleaved (i.e., Experiment 4B) at test, with different participants taking part in each experiment. Both versions were based on the same within-subjects, fully factorial design with cue condition (4) and exemplar type (2) as factors. The four levels to the cue condition were: shape only (S_v); shape with visual motion (S_vM_v); shape with tactile motion (S_vM_t) and shape with both visual and tactile motion (S_vM_{vt}). The exemplar factor had two levels: learned or novel. Our primary outcome measure of interest in the test phase was categorisation accuracy.

Procedure

The experiment was built using PsychoPy (Pierce et al., 2019; Pierce et al., 2022, v2024.2.3) and delivered online through Pavlovia (<https://pavlovia.org/>) on Android mobile devices. To mask the sound of the vibration stimuli, continuous Brown noise was presented throughout the study at a volume level determined based on an initial method of adjustment procedure. In this procedure, participants were presented with a repetitive vibration stimulus of 200 ms duration with a 500 ms ISI, alongside continuous Brown noise. They were asked to adjust their phone volume until they could no longer hear the sound of the vibration. Following this, the participants were asked to not adjust the volume on their phone for the duration of the entire experiment.

Following the method of adjustment procedure participants began the learning phase (Figure 25A). Each trial in the learning started with a 250ms fixation cross, followed by 2 seconds the presentation of the object stimulus in which three cues were combined as a visually moving object presented in synchrony with tactile vibrations. The stimulus was followed by a 250ms visual mask.

The participant was then presented with a screen indicating the response options (i.e., 'Category A' or 'Category B'). A response triggered the offset of the response screen, if a response was not made in 4 seconds the task progressed automatically. Feedback on both accuracy and response time was presented for 750 ms at the end of each learning trial, e.g., green 'Correct' or red 'Incorrect' and 'Too slow! Please respond faster' respectively). During the learning session, all objects were displayed with visual and (correlated) tactile motion (S_vM_{vt}); therefore, category membership was defined by all three cues equally. Each shape was associated with one of 4 visual motion patterns (swing, jump, roll, or shake), which were randomly assigned to the shapes across participants (but the shape-motion associations remained constant for each participant). The group of shapes assigned to each category were counterbalanced across participants. An accuracy threshold of 75% was required at the end of the learning session to continue to the categorisation test. If participants failed to reach this threshold within 3 repetitions of the trials, the study was terminated. There was a maximum of 54 trials (18 object shapes, with no more than 3 repetitions) during learning. Trial order was fully randomised across participants.

In the test phase, 28 stimuli (14 per category) were presented, including the 9 previously learned and 5 novel exemplars per category. Each test trial began with the presentation of a 250ms fixation cross, followed by an image of a stimulus (depending on cue condition) alongside response options, participants were given 4 seconds maximum to make a response before the task progressed automatically. A visual mask appeared for 250 ms after the cue to indicate the end of each trial. No feedback was provided at test. Each stimulus was presented under one of the four cue conditions (S_v , S_vM_v , S_vM_t , S_vM_{vt}) and participants were instructed to categorise each stimulus as either 'A' or 'B' as accurately and quickly as possible. In Experiment 4A, cue condition was blocked at test, such that one block included trials in which stimuli were displayed from one cue condition only (e.g., S_v only). The blocks were presented in a random order across participants. In Experiment 4B all trials were presented as interleaved and in a random order across

participants. For both Experiment 4A and 4B, no feedback was provided during the categorisation test. Each participant took approximately 18 minutes to complete the experiment.

Analysis

The data were analysed with R via RStudio (R Core Team, 2021). To assess the effect of cue condition and exemplar learning on categorisation performance, we fitted a generalised linear mixed-effects model (GLMM) on the raw participants' responses (0, 1) with a binomial distribution and logit link function. The model included cue condition (S_v , S_vM_v , S_vM_t , S_vM_{vt}), category exemplar (learned or novel), and their interaction as the predictors of interest. To establish statistical significance, likelihood ratio tests (type II) were performed to compare the fit of the model with and without that predictor of interest, the model fit was evaluated via AIC, BIC, and log-likelihood values. 'Participant' was included as a random intercept, and a random slope for 'exemplar' was also included to account for within-subject variability in the effect of exemplar on categorisation performance¹. Reaction times were not analysed as a primary measure due to the differential nature of the stimuli across cue conditions (for example, full motion duration of 2 seconds), which likely introduced systematic differences unrelated to categorisation difficulty (see Supplementary Materials – S16).

Models were fitted using the 'lme4' package (Bates et al., 2014). Post-hoc comparisons were conducted using the 'emmeans' package (Lenth Russell et al., 2022) and the Bonferroni correction was applied to correct for multiple comparisons (in which case the corrected p values are reported). Fixed effects from the final model were converted from log-odds to predicted probabilities using the inverse logit transformation to facilitate interpretation (Muller & MacLehose, 2014). Experiment 4A was pre-registered prior to data collection on the Open Science Framework, <https://osf.io/2ms48> and further details for Experiment 4, including the analysis

¹ Note: The structure of the random effect changes in the experiments, due to conversion issues. Experiment 4(A) has 'task' and 'exemplar', 4(B) has 'exemplar', 5(A) has 'task' and 5(B) has none, and Experiment 6 has 'exemplar' as random intercept.

script, are also available at https://osf.io/s369c/?view_only=9015e8935be24628b5b15596b8eb6271.

4.4.2 Result

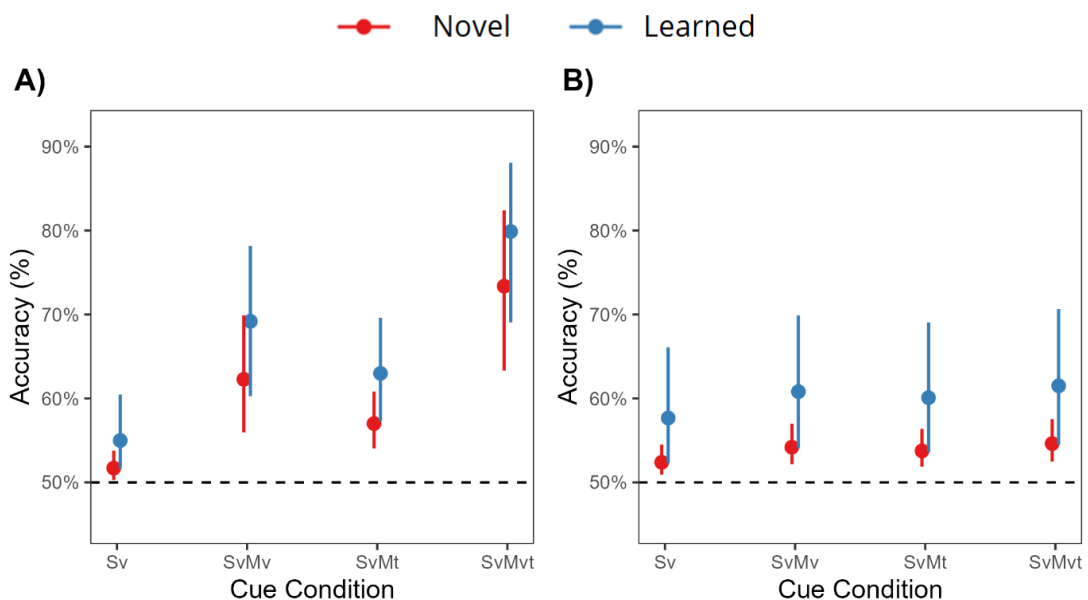
The mean categorisation performance across each of the cue conditions and exemplar types is presented in Figure 26 for the blocked (A) and interleaved (B) versions, respectively. In Experiment 4A, in which trials were blocked at test, likelihood ratio tests indicated statistically significant main effects of cue condition, $\chi^2(3) = 26.5$, $p < .001$, and exemplar type, $\chi^2(1) = 11.8$, $p < .001$, on categorisation accuracy. However, the cue condition by exemplar type interaction did not significantly contribute to the model, $\chi^2(3) = 6.4$, $p = .094$. Post hoc comparisons for the main effect of cue condition confirmed that categorisation accuracy was significantly higher for the S_vM_v (odds ratio, $OR = 0.395$, 95% CI [0.224, 0.698], $p = .0001$), S_vM_t ($OR = 0.565$, 95% CI [0.376, 0.848], $p = .0013$), and S_vM_{vt} ($OR = 0.209$, 95% CI [0.101, 0.430], $p < .0001$) compared to the S_v condition. Moreover, categorisation accuracy was significantly lower in both the S_vM_v ($OR = 0.528$, 95% CI [0.330, 0.845], $p = .002$) and S_vM_t ($OR = 0.370$, 95% CI [0.208, 0.657], $p < .0001$) conditions, compared to the S_vM_{vt} condition, as shown in Figure 26A. No significant differences were observed between the S_vM_v and S_vM_t conditions ($OR = 1.429$, 95% CI [0.871, 2.344], $p = .3426$). Post-hoc comparisons on the main effect of exemplar type suggested that participants' categorisation performance was less accurate for the novel (75%) compared to the learned (82%) object exemplars ($OR = 0.67$, 95% CI [0.538, 0.834], $p < .001$; see Figure 26A).

Likelihood ratio tests on the performance in the interleaved version of the Experiment 4B suggested similar effects. We found statistically significant main effects of cue condition, $\chi^2(3) = 12.2$, $p = .007$, and exemplar type, $\chi^2(1) = 8.08$, $p = 0.005$, in the model predicting categorisation accuracy. Again, the cue condition by exemplar type interaction did not significantly contribute to this model, $\chi^2(3) = 1.47$, $p = .69$. Post hoc comparisons for the cue condition effect revealed that accuracy was significantly lower in the S_v condition compared to the S_vM_v ($OR = 0.806$, 95% CI

[0.654, 0.993], $p = .039$) and S_vM_{vt} conditions (OR = 0.772, 95% CI [0.627, 0.950], $p = .006$), as shown in Figure 26B. However, we failed to find a significant difference between the S_v and the S_vM_t condition (OR = 0.845, 95% CI [0.686, 1.040], $p = .194$). Additionally, no significant differences were found between the two-cue conditions (S_vM_v vs S_vM_t : OR = 1.05, $p = 1$); and there was no benefit found to the S_vM_{vt} condition relative to either the S_vM_v (OR = 0.958, $p = 1$) or S_vM_t conditions (OR = 0.914, $p = 1$). Regarding the exemplar type, categorisation accuracy was significantly lower for novel (65%) compared to learned (76%) exemplars (OR = 0.597, 95% CI [0.422, 0.844], $p = .004$), as shown in Figure 26B. Finally, a one-sample t-test showed that accuracy in the S_v condition was significantly above chance for both Experiment 4A ($M = 0.61$, $SD = 0.23$, $t(44) = 3.21$, $p = .002$), and 4B ($M = .63$, $SD = .18$, $t(30) = 4.09$, $p < .001$).

Figure 26

Plots showing estimated mean categorisation accuracy predicted from the binomial mixed-effects model for the different cue conditions and for each exemplar type in Experiment 4.



Note: Results are shown when trials were (A) blocked or (B) interleaved. Error bars represent 95% confidence intervals. Dashed line indicates chance level (50%).

4.4.3 Discussion

Experiments 4A and 4B indicated a benefit of having all learned cues - object-shape, visual motion, and tactile vibrations - available at test, relative to when only the shape of the object exemplars was available. Categorisation performance was least accurate, although above chance, in the shape-only cue condition, suggesting that shape similarity alone did not result in robust categorisation. Performance to novel exemplars was further affected, although our data suggested that generalisation occurred even in the shape-only cue condition. Interestingly, while categorisation performance was affected by the number of available cues, the data suggest no specific benefit for any one sensory modality, at least when trials were blocked. In other words, performance across both two-cue conditions (S_vM_v and S_vM_t) was equally accurate, suggesting that visual and tactile motion contributed to category formation in a similar way. However, this result was influenced by the randomization procedure: in the interleaved trials tactile motion alone (S_vM_t) did not improve performance relative to the object shape-only cue (S_v), unlike in blocked conditions where both tactile (S_vM_t) and visual (S_vM_v) motion independently improved performance relative to static conditions.

In contrast, we found a stronger advantage for cue combination in the blocked (Experiment 4A) trial presentation. Under blocked presentations, there was a greater distinction in performance across the different cue conditions. This finding suggests that task consistency may support more effective use of the cues, whereas interleaving trials may add uncertainty reducing the observed benefit of information from combined cues on categorisation.

As expected, accuracy was consistently better for learned relative to novel exemplars, and there was no influence of trial presentation on the categorisation benefit for learned exemplars (7% and 11% difference between learned and novel exemplars in the blocked and interleaved versions, respectively), suggesting little effect of task demands. As noted earlier, exemplars that are positioned closer to the category boundary are typically more difficult to categorise than those that are positioned further from the boundary (Newell & Bulthoff, 2002). Therefore, the difference

in performance between learned and novel exemplars may be influenced by stimulus position within the category shape space as well as exemplar novelty. Future research could help tease apart the relative contributions of exemplar position in shape space and frequency of exposure on categorisation. Interestingly, the effect of cues did not differ between learned and novel exemplars in either version of the experiment, suggesting that the benefit afforded by multisensory motion cues is no higher for novel compared to learned exemplars.

The number of participant data sets included in the analyses was lower than the initial number of participants recruited. This was necessary for a number of reasons. First, we required that the participants first reach a learning rate of 75%, which is a reasonable rate given the number of factors involved in the experiment and is consistent with other studies on categorisation. Second, we also required that participants' performance was greater than chance level across the task. A number of participants failed to meet this requirement although they had successfully learned to categorise the objects. Although our overall attrition rate from initial recruitment appears high, it is not inconsistent with other studies of rates of successful categorisation performance in the lab (e.g. Smith et al., 2014; Roark & Chandrasekaran, 2023) or indeed in reports of drop outs from online testing in behavioural studies (e.g. Peer et al., 2022; although notably, data collected via Prolific is often considered of high quality).

Together, these findings suggest that a combination of cues from different sensory modalities can facilitate object categorisation and generalisation relative to object shape alone. Given that the visual and tactile motion cues to category membership were few in number and consistently reliable, our result that tactile motion did not benefit categorisation relative to shape-only cues in the interleaved trials might be considered surprising. Because of their consistency, both the visual motion and tactile motion cues might be expected to dominate categorisation performance relative to the shape cue alone. Since categorisation performance was better than chance to the shape cue (S_v) alone we can infer that shape was informative for categorisation. Indeed, the presence of each motion cue during test appears to have influenced categorisation

performance in an additive manner, at least when trials were blocked. Given this, in Experiment 5, we aimed to further explore the contribution of the shape cue by manipulating its informativeness relative to other cues on categorisation.

4.5 Experiment 5

Perceptual similarity may support the formation of categories, such that similar shapes are grouped together under a single category label (e.g., Nosofsky, 1986; Goldstone, 1994), even if encoded through touch (Cooke et al., 2007). However, in the real world, information about an object's shape can be affected by environmental factors such as distance, occlusion, lighting or viewpoint (e.g., Newell, 1998), thus reducing the reliability of shape similarity for determining categorisation. Consequently, other cues to categorisation may become salient such as the manner in which an object moves (Newell et al., 2004; Blake & Shiffrar, 2007). To investigate this, we reduced the informativeness of the object-shape similarity as a cue on category membership. Importantly, we maintained both the visual and tactile motion cues as fully predictive of category membership, as in Experiment 4. Consequently, we hypothesised that participants' categorisation performance, to both learned and novel exemplars, should be mainly influenced by motion cues. As in Experiment 4, we are testing this in both blocked (Experiment 5A) and interleaved (Experiment 5B) test conditions.

4.5.1 Method

Participants

An a priori power analysis was performed (PANGAEA (v0.2) for a within-subjects design) and the advised minimum sample size for 80% power to detect the effects of interest with size (Cohen's d) of 0.3, was 32 participants for each of the version of the experiment. All participants were recruited using Prolific (<https://www.prolific.co/>), and inclusion criteria were fluency in English, normal or corrected-to-normal vision and no hearing impairments. We initially recruited 85 participants, of whom 76 (i.e. 89%) successfully learned the categories. Of the learners, 10

participants failed to score above chance in the categorisation test and their data were not included in subsequent analyses. Following these exclusions, a total of 66 participants (mean age = 39.72 years, SD = 10.1; 48.5 % female) completed Experiment 5, with 33 assigned to the blocked version (Experiment 5A) and 33 to the interleaved version (Experiment 5B). All participants were naïve to the purpose of the study and received compensation to take part in the experiment (at a rate of £9.00 per hour). The study was approved by the Trinity College Dublin, School of Psychology Research Ethics Committee and complied with EU data protection legislation (GDPR).

Stimuli

The same stimulus set, and motion types described in Experiment 4 were used. However, we used a different arrangement of the object shapes across categories to reduce object shape similarity as a cue to category membership. To that end, we reassigned two out of the nine exemplars in each category to the opposite category. For example, referring to Figure 24, object shapes highlighted in red, which, in Experiment 4, were members of the same object category (e.g., Category A), were now assigned to the opposite category (e.g., Category B) during learning (and likewise for object shapes highlighted in green). Thus, only 78% of the neighbouring shapes were assigned to one category (see Figure 24) in the current experiment. The subset of exemplars that were novel but included in the test was the same as in Experiment 4 (and as shown in Figure 24).

Design and Procedure

Experiment 5 followed the same design and procedure as described in Experiment 4 with the same four cue conditions (S_v , S_vM_v , S_vM_t and S_vM_{vt}) and two exemplar levels (learned and novel). As in Experiment 4, trials were either blocked by cue (Experiment 5A) or interleaved (Experiment 5B) and participants were pseudo-randomly assigned to each version of the experiment.

4.5.2 Result

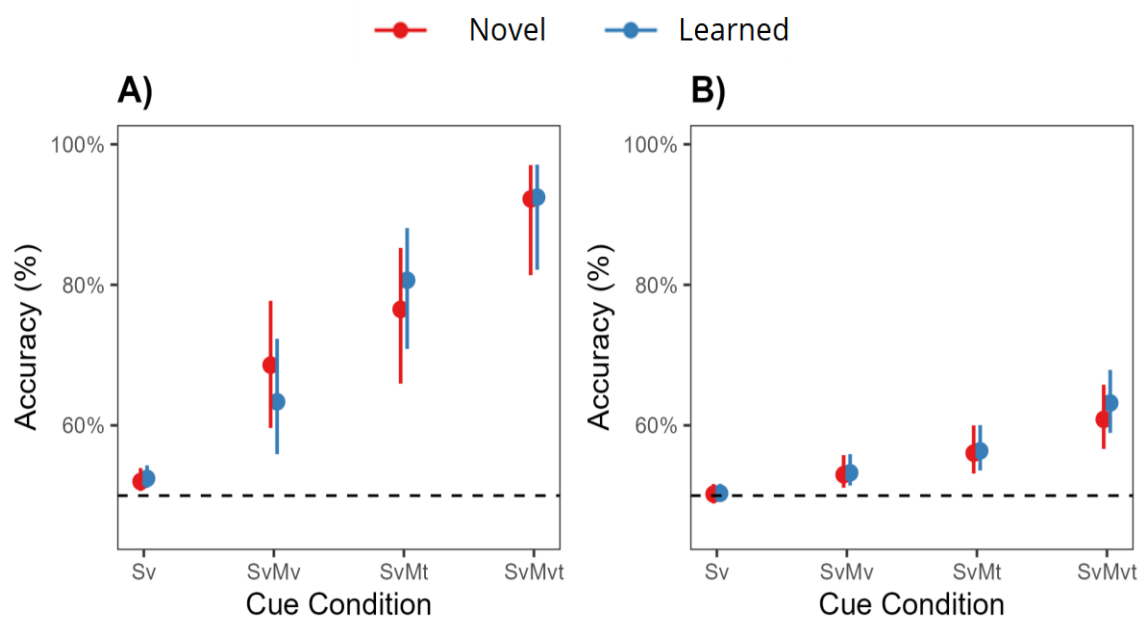
The main analysis procedure was identical to that of Experiment 4. Experiment 5 was pre-registered prior to data analysis on the Open Science Framework, <https://osf.io/dxwej> and further details of Experiment 5, including the analysis script, are available at https://osf.io/vkg7c/?view_only=08bc08d34e704ccbb9e90b1b4437f08b.

Categorisation performance across the four main cue conditions and each learned or novel exemplar condition is presented in Figure 27. First, categorisation performance on the blocked trials (Experiment 5A) was analysed using a likelihood ratio test, which compared the full model, which included the interaction between cue condition and exemplar type, to a reduced model without the interaction term. The comparison indicated that the full model significantly improved the fit of the data, $\chi^2(3) = 9.09$, $p = .028$, suggesting that the interaction between cue condition and exemplar type contributed to categorisation accuracy (see Figure 27A). There were significant main effects of both the cue condition and exemplar type on categorisation accuracy. Post hoc comparisons revealed that categorisation accuracy was significantly lower in the object-shape-only (S_v) condition compared to all the combined cue conditions. Specifically, accuracy to the S_v condition was significantly lower than to the S_vM_v condition (OR = 0.570, 95% CI [0.306, 1.064], $p = .018$), the S_vM_t condition (OR = 1.589, 95% CI [0.987, 2.557], $p = .013$), and the S_vM_{vt} condition (OR = 4.832, 95% CI [1.911, 12.222], $p < .001$). As in Experiment 4, no significant differences were found between the S_vM_v and S_vM_t conditions ($p = 1$), nor between each of these two-cue conditions and the three-cue condition (S_vM_v vs S_vM_{vt} , $p = 1$; S_vM_v vs S_vM_{vt} , $p = 1$). There was a main effect of exemplar type with better performance to learned (76%) over novel (65%) exemplars (OR = 0.597, 95% CI [0.422, 0.844], $p = .004$). The significant interaction between cue condition and exemplar type was driven mainly by performance between each of the two-cue conditions and the S_v condition (see Figure 27A): performance was better to the S_vM_v than S_v condition for the learned (OR = 0.570, $p = .018$), but not novel (OR = 0.770, $p = .282$) exemplars

and, similarly, performance was better to the S_vM_t than S_v condition for learned ($OR = 1.589$, $p = .013$) but not novel ($OR = 1.218$, $p = 0.282$) exemplars. Performance was better to the S_vM_{vt} condition than the S_v condition for the novel version only ($OR = 4.658$, $p < .0001$).

Figure 27

Estimated mean categorisation accuracy predicted from the binomial mixed-effects model for the different cue conditions and for each exemplar type in Experiment 5, in which shape similarity was a less predictive cue of category membership.



Note: The results are shown when trials were (A) blocked or (B) interleaved. Error bars represent 95% confidence intervals. Dashed line indicates chance level (50%).

We then conducted likelihood ratio tests on performance when trials were interleaved (Experiment 5B) which indicated a main effect of cue condition, $\chi^2(3) = 276.72$, $p < .001$, but not of the exemplar type, $\chi^2(1) = .953$, $p = .329$, in the model predicting categorisation accuracy, as shown in Figure 27B. Furthermore, the cue condition by exemplar type interaction did not significantly contribute to this model, $\chi^2(3) = 0.73$, $p = .866$. Post hoc comparisons indicated that categorisation accuracy was significantly lower in the S_v condition compared to the S_vM_v condition ($OR = 0.617$, $p < .001$), the S_vM_t condition ($OR = 0.455$, $p < .001$), as well as the S_vM_{vt} condition ($OR = 0.299$, $p < .001$). Moreover, performance was better to the S_vM_{vt} condition compared to the

S_vM_v ($OR = 0.484, p < .001$) and S_vM_t ($OR = 0.657, p < .001$) conditions. In contrast to the results from the blocked trials, performance to S_vM_v was lower than to the S_vM_t ($OR = 0.736, p = .0002$) condition when trials were interleaved.

Finally, a one-sample t-test showed that accuracy in the S_v condition was significantly above chance for Experiment 5A ($M = .61, SD = .17, t(32) = 4.27, p < .001$), but not for Experiment 5B ($M = .50, SD = .06, t(31) = -0.29, p = .770$).

4.5.3 Discussion

In Experiment 5, we reduced the informativeness of shape similarity as a cue to category membership to investigate whether participants would rely more on motion cues. The results suggest that when only the shape cue was available at test, performance was relatively poor. This was especially the case when trials were interleaved, i.e., when cue availability was less predictable from one trial to the next, relative to the version in which trials were blocked.

Categorisation performance was best when all cue combinations were available across experiment versions (blocked or interleaved). That is, the combination of shape with both motion cues led to significantly better accuracy than object-shape alone, for both learned and novel exemplars. However, in the blocked condition, this advantage did not extend to the two-cue combinations (i.e. S_vM_v or S_vM_t vs S_v independently) for novel exemplars, which did not significantly outperform the shape-only condition. The significant interaction between cue condition and exemplar type suggests that while all cue combinations supported learning of the trained exemplars, successful generalisation to novel exemplars relied more on the availability of both motion cues together. This suggests that when the reliability of shape as a cue to category membership is reduced, generalisation required the full cue-combination context at test.

Interestingly, when trials were interleaved, tactile motion provided a greater benefit than visual motion on both categorisation and generalisation performance. Again, this finding suggests a greater weighting of reliable cues, such as motion over shape. The relative benefit of tactile over

visual motion may be due to the easier integration of visual motion with, and therefore segregation of tactile motion from, the object shape (see, e.g., Chen & Spence, 2017).

In Experiment 4 and 5, the shape cue was always presented with motion cues during learning; therefore, it was unclear to what extent shape itself influenced learning and consequent categorisation and generalisation performance at test. Indeed, performance was consistently worse to the shape-only cue condition at test, and we attribute this performance to the absence of the other learned cues. In the following control experiment, we investigated the specific contribution of visual object-shape as a cue to category learning.

4.6 Experiment 6

In Experiment 6, participants learned categories defined solely by object shape (static images) and at test were assessed across the same four cue conditions used previously (i.e., S_v , S_vM_v , S_vM_t , and S_vM_{vt}). The use of motion cues at test further allowed us to assess the added value of multisensory motion cues beyond shape information on categorisation and generalisation.

4.6.1 Method

Participants

We initially recruited 45 participants online using Prolific, 8 of whom failed to reach learning criteria and a further 7 failed to perform above chance at test and their data were not included in the analyses. Therefore, total of 30 participants (mean age = 37.44 years, SD = 10.1; 48.5 % female) successfully completed Experiment 6; the reimbursement rate and inclusion criteria were identical to those of the previous Experiments.

Stimuli

All the stimuli, and arrangement of shapes into the two categories, were identical to that of Experiment 4 (see Figure 24).

Design and Procedure

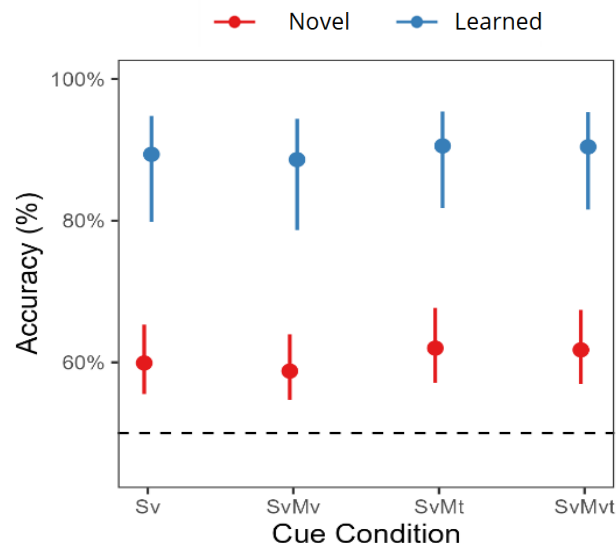
The experiment was based on the same design and procedure as described in Experiment 4 with the exception that the category learning task was performed on static object-shape stimuli. Also, we blocked trials by cue condition during the test session (as in Experiment 4A). Both learned and novel exemplars were presented at test in a random order across participants. Overall, the experiment took approximately 16 minutes for each participant to complete.

4.6.2 Result

The main analysis procedure to Experiment 6 was pre-registered prior to data analysis on the Open Science Framework, <https://osf.io/pem72> and further details, including the analysis script, are also available at https://osf.io/cnmhv/?view_only=eb8ecf26e5da471aa39b455a101b0dbb. Likelihood ratio tests indicated statistically significant main effects of exemplar type, $\chi^2(1) = 34.95$, $p < .001$, but not of cue condition, $\chi^2(3) = 5.25$, $p = .15$, to the model predicting categorisation accuracy. The cue condition by exemplar type interaction did not significantly contribute to this model, $\chi^2(3) = 2.748$, $p = 0.43$. Overall, irrespective of condition, participants categorisation performance was less accurate to the novel (66%) compared to the learned (86%) exemplars (OR = -7.38, 95% CI [0.08, 0.24], $p < .001$).

Figure 28

Estimated mean categorisation accuracy predicted from the binomial mixed-effects model for the different cue conditions and for each exemplar type in Experiment 6.



Note: Error bars represent 95% confidence intervals. Dashed line indicates chance level (50%).

4.6.3 Discussion

The results suggest that object shape was sufficient for category formation. However, we found that generalisation to novel exemplars was poor, with a drop of approximately 20% accuracy compared to the categorisation of learned exemplars (see Figure 28). Overall, these findings underline the important role of multiple cues during category formation and suggest that the benefit of visual and tactile object motion cues observed in the test session in the previous experiments are contingent on their prior integration into the category structure.

4.7 Exploratory Analyses of the Category Boundary

In addition to the main analysis in Experiments 4, 5 and 6, we conducted separate exploratory analyses to investigate whether cue condition also influenced categorisation performance for items that were positioned at the category boundary. The category boundary

was arbitrarily defined prior to learning as being at the half-way point in the shape space (i.e. between object stimulus numbers 14 and 15) such that objects positioned on one side of this boundary are members of 'Category A' and objects positioned on the other side are members of 'Category B' (see Figure 24), with category labels counterbalanced across participants. In these analyses, our primary interest was to examine the effect of removing cues relative to the learning condition (3 cues, $S_v M_{vt}$) and to determine whether this affected the slope of responses across the category boundary, and if so to what extent. A larger (i.e. steeper) slope value is indicative of a higher cross-category distinction (where the probability of a 'Category A' response is high for 'Category A' object members but low for 'Category B' object members), while smaller (i.e. shallower) slope value indicates a lower cross-category distinction (where the probability of a 'Category A' response is similar for 'Category A' and 'Category B' stimuli; see Chapter 2).

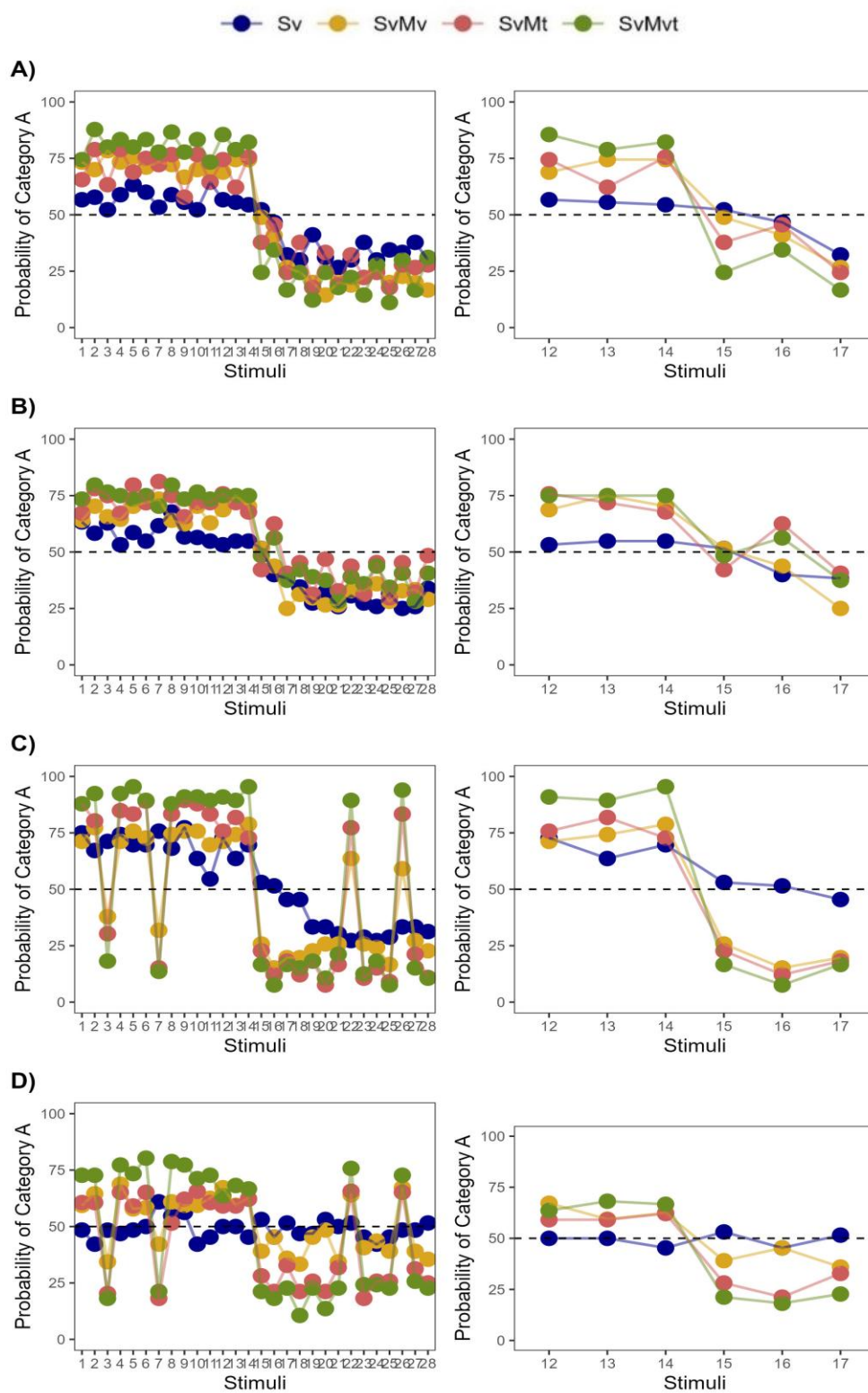
Analysis

To address this, we calculated the probability of a 'Category A' responses ($p(A)$) for the six exemplars closest to the category boundary, with 3 as members of 'Category A' (e.g., object stimulus 12, 13 and 14) and the other 3 as members of 'Category B' (e.g., 15, 16 and 17). We then took the average of each participant's accuracy responses across the 3 stimuli in each category and calculated the mean slope (slope= $p(A)_{\text{Category A}} - p(A)_{\text{Category B}}$) between the two categories, for each participant and cue condition. We then conducted a generalised linear mixed-effects model (GLMM) on the raw participants' responses (0, 1) with a binomial distribution and logit link function, with slope as the outcome variable and cue condition (S_v , $S_v M_v$, $S_v M_t$, $S_v M_{vt}$) as the predictor of interest. Participant was included as a random intercept.

4.7.1 Results

Figure 29

Plot showing the mean categorisation responses to objects positioned closest to the category boundary in Experiment 4 (A and B) and 5 (A and B).



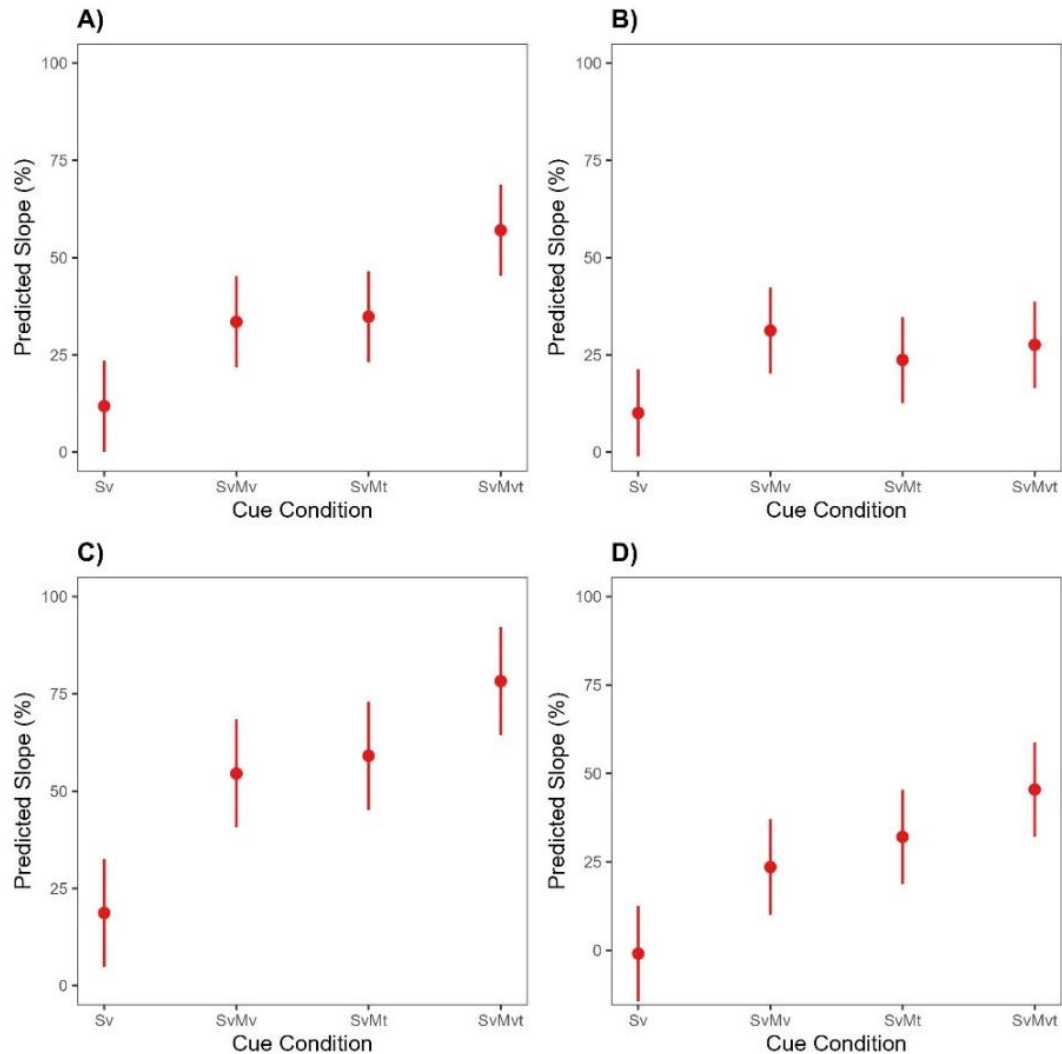
Note: The plots on the left panels depict the probability of ‘Category A’ responses to each of the stimuli across the cue conditions. The plots on the right panel provide a more detailed view of the slope between the category responses to three objects positioned each side of the category boundary and per each cue condition. Each row depicts responses from each experiment: Experiment 4A (panel A), Experiment 4B (panel B), Experiment 5A (panel C) and Experiment 5B (panel D). The different colours in the plots represent each of the cue conditions: blue, S_v ; yellow, S_vM_v ; red, S_vM_t ; green, S_vM_{vt} . The dashed horizontal line at 50% indicates the chance level of categorisation.

The probability of ‘Category A’ responses made to each of the object stimuli in the shape space, across the cue conditions, for the different experiments, are shown in Figure 24. In Experiment 4A, a likelihood ratio tests revealed a significant main effect of cue condition, $\chi^2(3) = 38.16$, $p < .001$ (see Figure 30A) on slope value. Post hoc comparisons showed that the slope to the S_v condition was significantly lower (i.e. shallower) than to either of the S_vM_v (estimated marginal means, EMM difference = -21.7, 95% CI [-40.1, -3.23], $p = .012$), S_vM_t (EMM = -23.0, 95% CI [-41.4, -4.52], $p = .006$) or S_vM_{vt} (EMM = -45.2, 95% CI [-63.6, -26.75], $p < .0001$) conditions. In addition, the slope values to both the S_vM_v (EMM = -23.5, 95% CI [-42.0, -5.08], $p = .005$) and S_vM_t (EMM = -22.2, 95% CI [-40.7, -3.78], $p = .009$) cue condition were significantly smaller (shallower) than to the S_vM_{vt} condition, whereas the slope values to the S_vM_v and S_vM_t did not differ ($p = 1$) from each other. The slope value to the three-cue condition was therefore the highest (steepest) than to any other cue combination. For the responses in Experiment 4B, a likelihood ratio test also revealed a main effect of cue condition, $\chi^2(3) = 14.17$, $p = .003$ (see Figure 30B). As in Experiment 4A, post hoc contrasts indicated that slope value to the S_v condition was significantly smaller than to the S_vM_v (EMM = -21.18, 95% CI [-37.04, -5.32], $p = .003$) and S_vM_{vt} (EMM = -17.53, 95% CI [-33.39, -1.67], $p = .022$) conditions. However, it was not significantly different from the slope value to the S_vM_t condition ($p = .136$). Also, there were no significant differences found between the slope values to any of the two cue conditions and the three cue condition (all $p = 1$).

In Experiment 5A, a likelihood ratio test again revealed a main effect of cue condition, $\chi^2(3) = 38.57$, $p < .0001$ (Figure 30C). Post hoc comparisons showed the slope value to the S_v condition was significantly shallower than to all other cue conditions: S_vM_v (EMM = -35.7, 95% CI [-60.1, -11.59], $p = .008$), S_vM_t (EMM = -40.4, 95% CI [-64.7, -16.13], $p < .001$), and S_vM_{vt} (EMM = -59.6, 95% CI [-83.9, -35.33], $p < .0001$). No significant difference was found between slope values to the two-cue and the three cue conditions: S_vM_v and S_vM_{vt} ($p = .059$), S_vM_t and S_vM_{vt} ($p = .214$), and S_vM_v and S_vM_t ($p = 1$). As in the other experiments, in Experiment 5B, a likelihood ratio test revealed a significant effect of cue condition, $\chi^2(3) = 27.55$, $p < .001$ (Figure 30D) with post hoc comparisons revealing a significantly lower slope to the S_v than all other conditions: S_vM_v (EMM = -24.5, 95% CI [-47.9, -1.05], $p = .036$), S_vM_t (EMM = -33.0, 95% CI [-56.3, -9.73], $p = .001$), and S_vM_{vt} (EMM = -46.4, 95% CI [-69.7, -23.11], $p < .0001$). No significant difference was observed between the two cues conditions and with three cues condition (all $p > .07$). For completeness, in Experiment 6, a likelihood ratio tests indicated no statistically significant effect of cue condition, $\chi^2(3) = 1.48$, $p = .69$, which was expected (see main results).

Figure 30

The plot shows the main effect of Cue Condition on the slope values to responses at the category boundary in Experiment 4 (A and B) and 5 (A and B).



Note: Experiment 4A (panel A), Experiment 4B (panel B), Experiment 5A (panel C) and Experiment 5B (panel D). A higher slope value indicates a steeper slope between category responses to object members of 'Category A' and 'Category B'. Error bars indicate 95% confidence intervals.

4.7.2 Discussion

Across three experiments, participants' categorisation performance on the boundary was affected by the removal of cues relative to the three-cue learning condition (SvMvt). Slope values

show participants' sensitivity to the category boundary: steeper slopes indicate sharper discrimination, while shallower slopes indicate weaker or noisier category use. Results showed significant main effects of cue condition on slope values in all but Experiment 6, suggesting that reducing the available cues weakened sensitivity to category distinctions. Specifically, the shape-only (S_v) condition consistently led to the shallowest slopes, suggesting that participants did not rely strongly on shape alone, likely because the learned category was defined by three cues, thus removing cues might have weakened category retrieval. The two-cue conditions (S_vM_v , S_vM_t) led to intermediate slopes, suggesting partial retention of category discrimination when one cue was removed. Finally, the three-cue condition (S_vM_vt) led to the steepest slopes, suggesting maximal sensitivity when all cues were available. In Experiment 6, as expected, removing cues had no effect.

4.8 General Discussion

Across three categorisation experiments, delivered via mobile phones, we examined whether previously learned multisensory categories, defined by shape, visual motion, and tactile vibration – could support object categorisation and generalisation. The motion cues were correlated across modalities and predictive of category membership (Experiment 4 and 5). Although shape similarity underpinned category membership in Experiment 4, its reliability was reduced in Experiment 5. Overall, the results of both experiments suggested that visual and tactile motion facilitated both the categorisation and generalisation of exemplars relative to shape-only information. The results of Experiment 6 confirmed that visual shape alone could support category learning in our tasks, although generalisation was poor. In addition, the task format, i.e. whether trials were blocked by cue or interleaved, significantly modulated the observed multisensory benefits: whereas blocked trials (Experiments 4A and 5A) consistently led to higher accuracy and better generalisation, an interleaved presentation (Experiments 4B and 5B) resulted in smaller performance differences between cue conditions and a less robust contribution from tactile cues. Our findings elucidate

the featural and context-sensitive nature of multisensory categorisation and demonstrate that the effectiveness of cross sensory cues for categorisation is dependent on cue reliability as well as task structure.

Our results align with previous evidence on the role of cue reliability on perception (Oruç et al., 2003; Helbig et al., 2012; Bankieris et al., 2017) whilst extending this to the realm of category learning. For example, when the inter-object similarity of shapes was reduced as a cue to categorisation (Experiment 5), and motion cues remained fully informative, participants' categorisation performance was influenced more by the tactile and visual motion cues, particularly under blocked conditions. Our results further support the idea that tactile cues are not merely supplementary but can be integral to the formation of category representations (Yildirim & Jacobs, 2013; Gaissert & Wallraven, 2012), especially when they are temporally aligned and task-relevant (Nordmark, 1978; Heller, 1989). Indeed, tactile motion cues played an important role in enhancing categorisation performance in all experiments, suggesting that individuals can strategically adapt their category learning performance to prioritise diagnostic information. Even under interleaved trial presentation (Experiment 5B), performance in combined cue condition remained high, suggesting that redundant motion cues can compensate against uncertainty from the shape cue.

The different effect of motion cues under blocked or interleaved trials suggests that task demands can influence cue combination, including across modalities (Carvalho & Goldstone, 2019; Maddox et al., 2006). Although some evidence suggests that interleaved presentation enhances generalisation due to discriminative contrast (Brunmair & Richter, 2019; Kornell & Bjork, 2008; Carvalho & Goldstone, 2014; 2015), including across modalities (Abel, 2023), we found that categorisation performance was weaker for novel than for learned exemplars. Although the reason for this inconsistency is unclear, we suggest that interleaved presentation may either trigger a greater reliance on individual cues or modalities or may impair the retrieval of specific

cue-category associations. Under blocked conditions, the motion cues integrated into the object representations may act as contextual supports in memory for categorising the object shape (Duarte et al., 2023, 2025).

In Experiment 6, the substantial (~20%) drop in performance to novel exemplars relative to the performance in previous experiments supports the idea that multisensory encoding during learning, rather than test-phase cue availability, is critical for supporting generalisation. This is consistent with evidence that learned multisensory representations subsequently support better generalisation across modalities (Yildirim & Jacobs, 2013). Interestingly, there was no benefit for motion cues at test suggesting that unfamiliar—but potentially informative, cues are not automatically used without prior association.

Moreover, considering the exploratory analysis conducted on the category boundary, we observed that removing cues compared to the three-cue learning condition (S_vM_{vt}) significantly reduced participants' categorisation sensitivity, as reflected in shallower slopes. The shape-only (S_v) condition produced the lowest slopes, indicating minimal reliance on shape alone, while the two-cue conditions (S_vM_v , S_vM_t) yielded intermediate slopes, showing partial retention of category discrimination when one cue was removed. This pattern suggests that category representations rely on the combination of multiple cues, and removing any component weakens retrieval and discrimination, supporting the role of cue combination in learned categories.

Our ability to categorise objects in the real world is likely to be dependent on a complex interaction between the nature and type of sensory information available, prior knowledge, and other context-dependent factors related to the task. For example, only tactile motion seems to have had a differential influence on categorisation and generalisation performance across task structure (interleaved or blocked trials). Moreover, methodological factors may help explain some of the inconsistencies in the reported findings in the literature (Wang & Zeng, 2022; Roark & Holt, 2022; Roark et al., 2021; O'Dowd et al., 2025). Consistent with this complexity, our findings suggest

that no single mechanism fully explains our data. In an attempt to investigate whether the observed cue combination benefits reflect multisensory integration (Shams & Seitz, 2008) or if categorisation responses go beyond predictions based on the summation of probability alone (see Nardini et al., 2010), we performed supplementary analyses (see Supplementary Materials-S12) on generalisation performance in Experiments 4 and 5, across presentation format. Results suggested that participants reweighted cues, relying more on motion (visual and tactile) than shape, consistent with reliability-weighting models of integration (e.g., Alais & Burr, 2004; Körding et al., 2007). Future studies are needed to investigate precisely how cue combination benefits learning and subsequent performance. In general, our results support the use of flexible, context-dependent integration strategies for learning novel object categories, and therefore help contribute to the ongoing debate about the role of multisensory information in object categorisation (Newell et al., 2023).

Our task was designed to enable remote data collection online via mobile devices. Although object perception experiments involving touch typically necessitate in-person testing (see e.g. Newell, 2004, 2010; Woods & Newell, 2004), technological improvements have meant that remote psychophysical testing using different modalities is rapidly increasing (e.g. Marin-Campos et al., 2021; Zhao et al., 2022; Hirst et al., 2025; Inuggi et al. 2024). The benefits for remote assessment are well-documented including the recruitment of larger and more heterogeneous participant samples (Grootswagers, 2020), and easier longitudinal follow ups. In particular, our work has important practical implications for the remote testing of multisensory category formation. Recent research from our team has validated the use of vibrations for remote testing contexts (Hirst et al., 2025), however to our knowledge no study has yet used this method to deliver complex patterns of vibration stimuli. This is therefore the first study to use complex vibration stimuli, delivered via the browser, to explore multisensory processes. It is of practical note that this method was successful, participants identified the correspondence between vibration patterns and visual stimuli with ease (Supplementary material – Table S7) and this

enabled us to conduct the current design and share this method for future research. We argue that remote delivery of tactile stimulation through mobile devices is an ecologically valid assessment method given that tactile feedback is increasingly integrated into technology (Muender et al., 2022; Ziat et al., 2022; Zwoliński et al., 2022). Indeed, technological advances now mean that many everyday objects incorporate some form of vibration feedback: you may feel resistance in the steering wheel of your car when changing lanes, the driver's seat might pulsate to signal danger, or your mobile phone may vibrate to signal an incoming call. Tactile feedback has also been shown to be effective at improving task performance (Fang et al., 2023) and the sense of presence (Huard et al., 2022; Kim et al., 2022). Moreover, delivery through commodity devices has real world implication for app design (in gaming or training). Despite these potential benefits, few published studies have utilised vibration delivery through mobile phones to study perception, and those that have report similar effects as lab-based studies (Inuggi et al., 2024; Hirst et al. 2025). Methodologically, the use of smartphone-delivered tactile cues offers an accessible and ecologically valid way for investigating multisensory interactions, particularly in applied domains such as education, human-computer interaction, and rehabilitation, where redundant sensory input, including from touch, can support cognitive abilities (see e.g. Gori et al., 2022).

4.9 Conclusions

Chapter 4 investigated whether visual and tactile motion cues enhance the categorisation of novel object shapes, and whether these multisensory categories support generalisation to new exemplars. Across three experiments delivered via mobile phones, we tested whether previously learned multisensory categories – defined by shape, visual motion, and tactile vibration – could support object categorisation and generalisation, across blocked versus interleaved cue-conditions at test. In Experiment 4 all cues were fully predictive of category membership, while in Experiment 5 the reliability of shape information for categorisation was reduced. In Experiment 6,

categories were learned with static shape only. When cues were equally reliable (Experiment 4), performance improved when all cues were available at test, and this was replicated when shape reliability was reduced (Experiment 5). Task format modulated these benefits: blocked presentation (Experiments 4A and 5B) yielded higher accuracy and stronger generalisation, whereas interleaved presentation (Experiments 4B and 5B) reduced performance differences and weakened contributions from tactile cues. Finally, visual and tactile motion facilitated categorisation and generalisation relative to shape-only learning (Experiment 6).

Collectively, these findings demonstrate that multisensory motion cues promote object category formation and generalisation, with effectiveness influenced by cue reliability and task structure. The results have important implications for our understanding of the underlying dynamic and multisensory nature of object categories and the predictive role of multisensory features on category formation. To our knowledge, this is also the first study to use complex vibration stimuli, delivered via the browser, to investigate multisensory processes. These results and methodology provide a valuable resource for future work on tactile and multisensory processing in applied contexts.

Chapter 5

**A crossmodal shape conflict results in a haptic over vision bias in the categorisation of novel 3D objects:
a virtual reality study**

5 A Crossmodal Shape Conflict Results in a Haptic Over Vision Bias in the Categorisation of Novel 3D Objects: a Virtual Reality Study

5.1 Abstract

Understanding how the brain resolves conflicting information from different sensory modalities is central to models of multisensory perception. While vision typically dominates in many perceptual domains, haptic input can provide richer spatial detail in three-dimensional object perception. This study examined how the perceptual system resolves visuo-haptic conflict in a categorisation task. Specifically, we tested whether categorisation varied with conflict magnitude and whether one modality became dominant under higher conflict. Using a shape-conflict paradigm, we conducted a free categorisation task with unfamiliar objects by parametrically manipulating the visual–haptic shape discrepancies (from small to large, across six steps) between three-dimensional object shapes. Participants actively explored the physical objects using their hands while, at the same time, viewing their visual counterparts in an immersive virtual reality (VR) environment. Results suggest that participants maintained stable category representations across all levels of conflict but that category decisions relied predominantly on haptic information. This finding suggests that categorisation was biased towards haptic input which may have provided more detailed shape information than the visual input via VR. Consistent categorisation performance despite sensory conflict implies access to shared, modality-independent representations rather than trial-by-trial cue reweighting. Limited realism and embodiment in VR may have further reduced the perceived consistency of visual information. These findings show that during visuo-haptic conflict, categorisation can flexibly adapt to the more stable cue, or task demands and stimulus properties, and have implications for VR design and models of multisensory integration.

5.2 Introduction

Everyday object recognition and categorisation often rely on the integration of information from multiple sensory modalities (Ghazanfar & Schroeder, 2006), particularly perceiving object shapes through vision and touch. Shape is proposed as a fundamental diagnostic feature for both object recognition (Marr, 1982; Biederman, 1987; Ullman, 1989; Tarr & Bülthoff, 1995) and categorisation (Katz & Fodor, 1963; Rosch & Mervis, 1975; Nosofsky, 1986) due to its constancy when other features change, for example colour or texture. Indeed, shape provides information about an object's underlying structural organisation and spatial configuration (Biederman, 1987; Lescroart, & Biederman, 2013) and these descriptions offer robust viewpoint-invariant shape representations (Tarr & Bülthoff, 1995), which in turn are essential for determining its identity (Ayzenberg & Lourenco, 2019). Moreover, humans, particularly children, develop a 'shape bias', prioritising an object's shape over other object properties when forming categories (Colunga & Smith, 2008; Vlach, 2016), supporting the role of shape as a favoured dimension for organizing knowledge about objects.

As key dimension for both object recognition and categorisation, object shape can be inferred both by vision and haptic information, which the brain optimally combines to create more reliable shape representations (Ernst & Bülthoff, 2004). The sense of touch allows us to instantly perceive objects and to combine multiple features into percepts that support categorisation. For example, while searching for a phone in a bag, one might grasp a pack of tissues of similar shape but quickly distinguish between the two based on spatial qualities perceived through touch. Unlike vision, which reconstructs a 3D representation from a 2D retinal image, haptic perception comprises active and direct exploration of objects in three dimensions (Klatzky & Lederman, 2011; Jansen et al., 2013), allowing for the extraction and process of geometric form and other unique material properties (e.g. texture, temperature) of objects (Lederman & Klatzky, 1987; Lederman & Klatzky, 2009; Jones & Lederman, 2006; Kappers & Bergmann Tiest, 2013; Masson et al., 2016).

Due to the active and direct nature of exploration, haptic perception it is often less vulnerable than vision to perceptual or environmental ambiguity, or spatial illusions (Wijntjes et al., 2009; Fairhurst et al., 2018, but also see Henriques & Soechting, 2003). For instance, in occlusion scenarios, e.g., reaching into a bag or manipulating objects in low light, haptic information is the main system used for extracting spatial and geometric structure of the objects (Lederman & Klatzky, 1987; Gentaz et al., 2008). Through haptic exploration, individuals can infer continuity across hidden surfaces (Takeichi et al., 2023), estimate global size and volume (Kappers & Bergmann Tiest, 2013; 2014), and reconstruct object identity despite poor or misleading visual input (Wijntjes et al., 2009). Overall, this indicates that touch enables object recognition and categorisation in different ways from vision.

The visual and haptic modalities provide partially overlapping but distinct information about object shape (Klatzky, & Lederman, 2010): while vision process global form, contour and colour information, haptic exploration provides access to detailed information about the texture, hardness, weight and curvature of objects through hand movements (Lederman & Klatzky, 1987). When visual and haptic shape information are congruent and available simultaneously, the brain can optimally combine these cues to form robust object representations (Ernst & Banks 2002), following the principles of reliability-weighted cue combination, in which each modality is weighted based on its relative reliability (Ernst & Banks, 2002; Ernst & Bühlhoff, 2004). This integration of object shape across modalities is supported by similar neural substrates; for instance, the lateral occipital complex (LOC), which was traditionally associated with visual object processing, but is also activated during haptic object exploration, suggesting a shared representation of object shape across modalities (Amedi et al., 2001; 2002; Stilla & Sathian, 2008). Moreover, the Intraparietal Sulcus (IPS) enables whole-from-part object reconstruction for both familiar and unfamiliar objects in vision and touch (Sathian et al., 2011; Kassuba et al., 2013; Lacey & Sathian, 2014) and serves as a key hub for object-directed motor actions for haptic exploration (Rice et al., 2006; Tunik et al., 2008). In addition, the anterior IPS (aIPS) involved in object size and

shape processing (Perini et al., 2020). Overall, these findings support the view that object shape is processed by the brain through the convergence of haptic and visual sensory information.

While vision and haptic modalities typically provide redundant or complementary information about objects (Newell et al., 2001; Ernst & Banks 2002; Burge et al., 2010; Takahashi & Watt, 2017), real-world scenarios frequently involve conflicting sensory inputs, such as situations in which the visual–haptic (VH) properties of an object do not align. A classic example is the size–weight illusion, in which larger objects of equal weight are perceived as lighter than smaller ones because visual expectations conflict with haptic information (Murray et al., 1999). Everyday experiences, such as mistaking a weighted blanket for a light wool one based on visual cues, similarly demonstrate how discrepancies between modalities are resolved through cue weighting, with haptic input typically guiding the final percept of weight. When sensory cues conflict across the visual and haptic modalities, the brain must resolve this discrepancy and integration processes might become more complex, leading to cue segregation and weighting or sensory dominance, depending on task demands, context, and cue reliability (Ernst & Banks, 2002; Welch & Warren, 1980). For example, cue dominance may emerge, whereby information from one modality systematically overrides the other for perception (Helbig & Ernst, 2007).

Among the factors modulating whether integration or segregation occurs, the assumption of unity—the belief that sensory signals originate from the same origin—constrains intersensory bias. For small discrepancies, unity is typically preserved, and perception may be a weighted compromise between the cues, whereas when discrepancies are large enough to imply distinct objects, unity breaks down, leading to segregation and reliance on the more reliable modality (Welch & Warren, 1980). Moreover, both task demands, and the reliability of the sensory input can modulate the relative weighting of the cues for perception. For instance, in tasks requiring high spatial resolution, visual information may dominate, whereas during texture discrimination task, haptic input may be more influential (Rock & Victor, 1964; Hollins et al., 2001; Ernst & Banks,

2002). In contrast, during texture or fine-grained shape discrimination tasks, haptic input may be more dominant (Heller, 1992; Hollins et al., 2001; Ernst & Banks, 2002; Jones et al., 2005). Importantly, dominance is flexible rather than absolute, and it can change depending on the reliability, and informational richness of the available sensory inputs, as well with context, attentional focus, and prior experience (Ernst & Banks, 2002; Gepstein et al., 2005). Furthermore, the degree of conflict also modulates perceptual outcomes: moderate incongruencies can increase uncertainty, slow down decision-making, and demand greater cognitive effort to reconcile competing cues (van Beers et al., 2002; Trommershäuser et al., 2011; Tang et al., 2025), while extreme conflicts may instead prompt greater reliance on a single modality, effectively ignoring the conflicting input (Colavita, 1974; Gepstein & Banks, 2003; Abtahi & Follmer, 2018).

In visuo-haptic object recognition, familiar and unfamiliar objects perception engages different neural mechanisms, with familiar objects relying on top-down visual imagery and memory retrieval, and unfamiliar objects processed mainly through bottom-up input (Lacey et al., 2009; Deshpande et al., 2009). This suggest that the processing of sensory information and influence of stored knowledge during object recognition may differ (Bulthoff & Newell, 2006). Regarding familiar objects, previous studies have shown that unimodal information, such as either vision (Peissig & Tarr, 2007; Grill-Spector, 2003) or touch (Gibson, 1962; Lederman & Klatzky, 1987; 1990;2004; Jones et al., 2005), is sufficient to aid learning and promote efficient object recognition (Newell et al., 2001; Lacey et al., 2009). Specifically, while visual object recognition is typically rapid and accurate (Thorpe et al., 1996), haptic recognition of familiar objects is also relatively fast and highly accurate (Lederman & Klatzky, 1985), with even a brief ‘haptic glance’ sufficient to recognise the objects (Klatzky & Lederman, 1995). While unimodal perception can be highly accurate in isolation, combined VH perception can lead to behavioural benefits, such as enhancing accuracy (Lacey & Sathian, 2014), target search (Wang et al., 2024), memory (Siqueira Rodrigues et al., 2024), and motor efficiency (Volcic & Camponogara, 2018), and faster responses times (Rowland et al., 2007) compared to unimodal conditions. However, haptic recognition can

be significantly more accurate in identifying complex geometrical shapes compared to visual-only and visuohaptic exploration (Jones et al., 2005). This suggests that while vision favours simple shapes, haptics' active exploration better supports detailed representations of complex, non-geometric forms. Regarding unfamiliar objects, information can be successfully transferred between vision and touch, leading to efficient cross-modal integration (Yildirim & Jacobs, 2013; Erdogan et al., 2014), as well category structure (Wallraven et al., 2013). However, this transfer can often be asymmetric: visual-to-haptic recognition of object sizes and shapes is often more accurate and faster compared to haptic-to-visual recognition (Easton et al., 1997; Lacey & Campbell, 2006; Lacey & Sathian, 2014; Newell et al., 2001). This asymmetry, which might be explained by modality-specific encoding strategies (Cooke et al., 2007), often involves differential cognitive costs (Lawson, 2009; 2014). Relative to unimodal object recognition, crossmodal recognition is often associated with a performance cost (Ernst et al., 2007; Casey & Newell, 2007; Lacey et al., 2007; Phillips et al., 2009; Lacey & Sathian, 2014), due to representational mismatches (Newell et al., 2001; Norman et al., 2008). For instance, Phillips et al., (2009) reported that while visual object discrimination remained accurate across objects of varying complexity, haptic and crossmodal recognition declined for more complex shapes, with crossmodal performance showing only limited benefit over vision alone for simpler objects. This suggests that while the brain can effectively transfer and integrate information across senses, information from a single-modality might be led to better performance in stimulus- and context-dependent way.

The interplay between visual and haptic cues is particularly important during object categorisation, where the brain must flexibly weigh competing inputs to assign objects to categories (Lederman & Klatzky, 1987; Adam & Noppeney, 2009; Jones et al., 2005; Pinotsis et al., 2019). When visual information is degraded, noisy, or semantically misleading, the contribution of haptics to the final percept increases (Abtahi & Follmer, 2018; List et al., 2014). Evidence has been reported for symmetric cross-modal transfer of category structure following visual- or haptic-only category learning, with performance improvements generalising to novel objects

across modalities (Wallraven et al., 2014). Thus, the resulting visuo-haptic shared representations may lead to the formation of more robust object categories than unisensory information alone (Yildirim & Jacobs, 2013; Gaissert & Wallraven, 2012; Haag, 2011; O'Callaghan et al., 2018; Broadbent et al., 2020). In situations in which the visual and haptic (VH) object information is conflicting, the perceptual response might depend on the assumed unity of sensory signals (e.g. if the information is from the same spatial location). Heller (1992) noted that in tasks requiring fine-grained geometric judgments, haptic information often biases final perceptual decisions under conflict. The degree of the conflict may be important in that if small, there may be no noticeable bias toward either modality on the final percept, however as these discrepancies increase perception might be biased towards the more task-appropriate modality, such as vision for spatial tasks, and haptics for material or fine-grained shape discrimination. Moreover, minor incongruencies can alter the perceived size or shape, with haptic feedback critically modulating or constraining visual-induced changes (Girondini et al., 2024).

5.3 Outline of the Experiment

The present study aims to investigate the interaction between vision and haptics in object categorisation, when the spatial information encoded by each modality is in conflict. While prior work has demonstrated how humans combine sensory cues for perception (Ernst & Banks, 2002), most studies have been conducted in constrained laboratory settings with limited ecological validity (Burge et al., 2010; Takahashi & Watt, 2017). Virtual reality (VR) provides a valuable alternative, allowing for high-fidelity rendering, real-time tracking of physical interactions (Sanchez-Vives & Slater, 2005; Sra & Schmandt, 2015), and active, naturalistic exploration of objects, an essential component of haptic perception (Gallace & Spence, 2014). The use of VR has already proven valuable in research on embodiment (Kiltner et al., 2012), sensory prediction (Tierl et al., 2015), perceptual training (Sanchez-Vives & Slater, 2005), and it is increasingly being applied

to research on multisensory integration (Marucci et al., 2021; Girondini et al., 2024; 2025; Woods et al., 2025).

In the present study, we employed a shape-conflict paradigm to assess the relative contribution of object shape encoded through touch and vision on the categorisation of unfamiliar objects. Using an immersive virtual reality environment (VRE), we systematically manipulated a conflict in object shape across the visual and haptic modalities. To that end, we used a parametrically morphed shape continuum of unfamiliar 3D objects which allowed for the precise control of the discrepancy between the seen and felt shapes. In the task, participants could freely categorise the 3D objects following active exploration. This setup enabled us to test whether categorisation responses vary with the magnitude of cross-modal conflict, leading to less stable categorisation with increasing conflict. Alternatively, increasing conflict may be associated systematic modality biases, such that a certain degree of conflict may result in information from one modality dominating the categorisation responses. In sum, the aim of the study was to elucidate the dynamic interplay between modalities during crossmodal conflict in an ecologically valid context.

5.4 Experiment 7

5.4.1 Method

Participants

All participants were recruited through the SONA system at Trinity College Dublin, based on the following inclusion criteria: fluency in English, normal or corrected-to-normal vision and no hearing impairments, no known tactile or neurological impairments. One participant reported high susceptibility to motion sickness or dizziness (assessed via pre-screening through the Motion Sickness Susceptibility Questionnaire MSSQ-Short; Golding, 1998) and therefore could not take part. An a priori power analysis was performed in PANGEA (v0.2) (Westfall, 2016) for a within-

subjects design. The advised minimum sample size for 80% power to detect the effects of interest with size (Cohen's d) of 0.3, was 12 participants. In total, and in accordance with design aspects of our experiment, 22 participants (mean age = 25.1 years, SD = 3.1; 23% female, 9% left-handed, 54.5% never used VR before) completed Experiment 7. All participants were naïve to the purpose of the study and were compensated at a rate of 10.00 Euro per hour. The study was approved by the Trinity College Dublin, School of Psychology Research Ethics Committee (approval number REF#2502) and complied with GDPR data protection legislation.

Stimuli and Apparatus

We created 36 novel object shapes (3D-designed) with uniform spacing between neighbouring objects (except for object shapes at each end of the shape space, as shown in Figure 31) for both visual and haptic exploration. The design was based on a morphing process described elsewhere (Li et al., 2020), and it was adapted from a circular shape space described in previous studies (Li et al., 2020; O'Dowd et al., 2025; see Chapters 3 and 4). We used 3D modelling software (Blender Foundation, 3.5.0, 2023, www.blender.org) to render each visual 3D object stimulus used in the present experiment. The 3D visual object stimuli were created using the following pipeline previously described in Chapters 3 and 4. We also created real versions of these objects by 3D-printing the shapes using a uniform white PETG. To stabilise the object shapes for haptic exploration, each object was positioned on a custom-built platform (a pyramid structure) as shown in Figure 31. The support structure allowed the objects to be positioned at a consistent height of 88mm and base dimension of 65x65mm, ensuring standardised visual and tactile presentation. For the purpose of the experiment, each (3D-printed) physical object used for haptic exploration was mounted onto a real table, at a standard distance from the participant. Furthermore, the position of the physical object was spatially aligned with its virtual counterpart both in spatial location within space as well as the viewpoint of the object. We also created four additional basic geometric shapes (e.g., sphere, cube, pyramid and cone) for use in a practice task

(using the same procedure for the main stimuli set). The full set of stimuli and rendering scripts are available via OSF (<https://osf.io/s369c/>).

Figure 31

An illustration of the object stimuli used in Experiment 7, arranged by stimuli set: Set A (panel A), Set B (panel B) and Set C (panel C), and practice stimuli (panel D). The images are of the real (3D printed) versions of the object stimuli arranged in a shape continuum from set A to set C.



We used a standalone head-mounted display (HMD; Meta Quest 3, Meta Platforms Inc.) to deliver an immersive virtual reality environment (VRE). This device allowed for precise presentation and manipulation of the three-dimensional object stimuli used in the experiment, with independently controlled visual and haptic shape features. The Meta Quest 3 features dual LCD displays (2064×2208 pixels per eye; a combined resolution of 4128×2208 pixels; $\sim 110^\circ$ horizontal field of view), has a built-in audio system, and supports a refresh rate of up to 120 Hz. The device provided six degrees of freedom (6DoF) tracking for both head and hand movements, allowing naturalistic object exploration without physical joystick controllers via integrated hand tracking. Haptic shape information was rendered using a physical proxy object (e.g., 3D printed

shape), synchronised with their visual counterparts in the Virtual environment (VE). Participants hands were displayed, in real time, as virtual, transparent images with a black outline within the VRE.

The experiment was run on a Dell Precision 3660 desktop workstation (Dell Technologies, USA) with a 12th Gen Intel® Core™ i9-12900K 3.20 GHz processor, 32 GB of RAM, and an NVIDIA GeForce RTX 3080 GPU (NVIDIA Corporation, USA) graphics card running the experimental software. The task environment and interactions were designed and implemented in Unity (Unity 6, version 6000.1.1f1) which projected to the headset. The experimental setup included two dedicated monitors for the experimenter. One monitor displayed the participant's VR scene in real time, enabling ongoing verification of visual–haptic alignment. The second monitor hosted the experimenter interface, which controlled trial onset and offset, displayed the list of upcoming virtual stimuli, and specified the corresponding haptic objects to be placed on the object holder. This dual-monitor arrangement allowed for a precise coordination of the virtual and physical stimuli. In order to match the virtual and physical space for each participant, a calibration trial was conducted at the start of each session. The calibration comprised of the following process: participants first lifted each joystick controller from its fixed holder on the desk, raised it vertically, and pressed the A (right joystick) and X (left joystick) buttons. This step allowed the headset tracking system to record the controllers' distance relative to a fixed reference marker positioned on the wall (see Figure 32). Participants then returned the controllers to their fixed holders and pressed the same buttons again. The headset tracking system used these two measurements to compute the relative position and orientation of the controllers, thereby aligning the virtual desk anchor with the physical desk surface. This procedure ensured the virtual plane matched the physical plane with precise accuracy.

A custom-made 3D-printed holder was used to position the physical stimuli. This holder—square, transparent, 65 (width) x 65 (length) x 10 mm (height)—was placed and glued to the centre

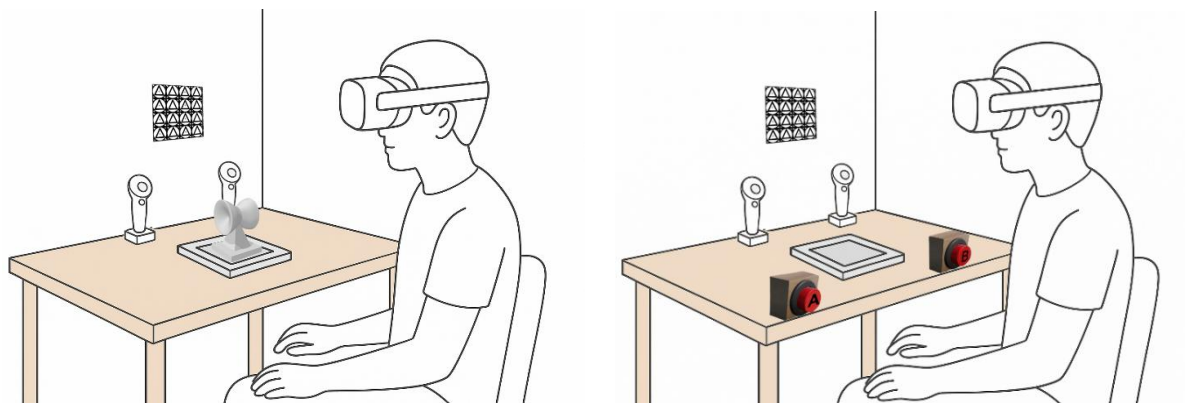
of the desk in a fixed location and always corresponded to the position of its virtual counterpart. This provided a consistent anchor for haptic stimuli and ensured that each virtual object was presented in precisely the same position as its physical equivalent. Participants were seated on a fixed chair and could reach forward with their dominant hand to interact with each physical object and its virtual counterpart which was positioned on a virtual table. Virtual response buttons were displayed within the VE and were activated by the participant with a press of their index finger. Details of the testing apparatus are shown in Figure 32.

Figure 32

A schematic drawing showing the task environment in Experiment 7. A) An illustration of the set up for the participant with the physical object presented in the centre of the table ready for exploration. B) The illustration shows the two (virtual) response buttons as would be viewed within the VE. These buttons allowed for the categorisation responses

A)

B)



Note: The image shows a participant wearing the VR headset seated at a desk with two joysticks fixed in their holders and the holder for the real object fixed at the centre of the desk.

This illustration was generate with assistance of AI (ChatGPT, <https://chat.openai.com>)

Design

For the purpose of our categorisation experiment, we divided the 36 object stimuli within the shape space (see Figure 31) into three sub-sets of 12 object stimuli each. Each participant was

randomly assigned to one of these objects sets and was therefore tested on that set alone, such that all object subsets were equally represented across the sample. Each participant was presented with stimuli from one set only across trials. Each of the 12 objects within the three stimulus sets was labelled according to its position in the shape space (1–12). For our design, and to ensure all stimuli were fully counterbalanced, the primary object shape was visual, and the comparison object shape was the real object (haptic).

Each trial involved a crossmodal pairing of a visual shape and a haptic shape (i.e. V-H). These V-H shape pairings were either congruent (i.e. the same shaped objects) or incongruent (i.e. a shape conflict). In the incongruent condition, we parametrically controlled the degree of shape conflict across modalities: this was achieved by choosing different visual and haptic objects that were separated by the number of neighbouring shapes on the shape space described above (see Figure 31). The crossmodal conflict increased from 2 to 6 separating shapes. For example, in Figure 31A, the visual object could be the first object from Set A (leftmost object) whereas the haptic object could be the fifth object (from the left) in Set A. We refer to these 'separating shape' numbers as 'steps' within the shape space to indicate the degree of V-H conflict. Accordingly, no V-H conflict, i.e. congruent V-H shapes, is described as '0-step' difference, and incongruent V-Hs are described as 2-, 3-, 4-, 5-, 6- steps, yielding a total of seven levels of conflict, without exceeding the stimuli set range (i.e., 1–12). We did not include '1-step' difference since the objects were considered too similar to each other in that pairing. Specifically, for haptic stimuli 1–10, visual paired stimuli were chosen by moving 2–6 steps rightward along the shape space shown in Figure 31, while for haptic stimuli 2–12, visual stimuli were chosen by moving 2–6 steps leftward along the shape space. This approach ensured that each haptic stimulus was paired with all possible visual stimuli within the within the 1–12 stimulus range for each set (see Figure 31) and led to 40 combinations for each pairing direction along the shape space (left- or rightward), resulting in a total of 80 incongruent trials per object set.

The experiment was structured around two blocks of trials in which the participant freely categorised each object by pressing one of two response keys within the VE (see Figure 32B). The position of these labelled (Category A or B) response keys was counterbalanced across participants. In the first block, we presented trials in which the V-H shape was congruent only (0-step difference) in order to provide a baseline measure of object categorisation from each participant. Each object was repeated 3 times in this block, with a total of 36 trials (3 repetitions of 12 objects, e.g. set A in Figure 31). The trials in the second block (Block 2) included both V-H congruent and V-H incongruent trials (2, 3, 4, 5 and 6 step differences). There were 24 V-H congruent trials and 80 V-H incongruent trials in the block, yielding a total of 104 unique visual–haptic shape combinations across trials in Block 2.

The order of the trials was randomised within blocks across all participants. Our primary measures of interest were categorisation choice ($p[A/B]$), and modality bias in the incongruent condition (proportion of responses aligned with visual versus haptic category). Furthermore, categorisation responses to trials in Block 2, in which objects lay across the individual category boundary (cross-category trials), were analysed separately to assess modality dominance and categorisation response bias. Additionally, we conducted exploratory analyses of categorisation response times (ms), although it is noted that these response times were constrained to the end of object exploration.

Procedure

The experiment was conducted in a research-dedicated testing laboratory which was dimly lit. After providing informed consent, participants completed a demographic questionnaire. Prior to the experiment, the VR headset was placed on the participant, and they completed a calibration phase (see Stimulus and Apparatus) to spatially align the virtual image with the position of the physical object (see Figure 32).

Following successful calibration of the spaces, the participant was instructed that their task was to (explore) each object and decide which of two categories (A or B) the object belonged. The participant was instructed to freely categorise each object into Category A or B based on their perception of its shape, both using visual and tactile information, by pressing one of two labelled buttons shown within the VE (see Figure 32). Participants were instructed to respond as quickly as possible and no feedback was provided. To familiarise the participants with the categorisation task, the experiment began with a practice phase (of 16 trials) using basic geometric shapes (e.g., sphere, cube, pyramid and cone). Once the practice was completed, participants proceeded to the main test.

In the main experiment, each trial began with an auditory 'start' cue (i.e., beep), which alerted the participant to start to explore the objects both visually and haptically using their dominant hand. The visual stimuli appeared in the centre of the virtual environment and its physical counterpart was positioned in front of the participant so that it was spatially aligned to the virtual object (see Figure 32). Each object stimulus was presented for a duration of 6000 ms. A 'stop' auditory cue then signalled the end of exploration, and the offset of the object presentation. Response times were measured from this offset. Self-timed breaks were provided between blocks on request.

Following the main task, participants completed four post-experiment questionnaires (SSQ - Freiwald et al., 2020; Kennedy et al., 1993; VEQ - Roth & Latoschik, 2020; IPQ - Schubert et al., 2001; and Realism Evaluation - Schmied-Kowarzik et al., 2023). Responses to these questionnaires helped to assess participants' subjective experience in the VR environment and their categorisation strategy (for detailed information refer to Supplementary Materials – Table S8). Participants took approximately 50 minutes on average to complete the whole experiment.

Analysis

We first ensured that we could rule out any differences in categorisation performance that were due to the specific stimuli sets used, such as whether individuals relied on different shape cues depending on the object set. These initial analyses revealed that there were no significant differences in categorisation responses across the object sets (for detailed information refer to Supplemental Materials, Figure S4). Once this was confirmed, and after data normalisation, we then performed the following analyses on the data. First, to analyse participants' category responses across blocks, we fitted logistic regression to each participant's responses for Block 1 (no-conflict) and Block 2 (conflict, congruent trials only). For each participant, the probability of responding 'A' ($P(A)$) was estimated as a function of stimulus position within the shape space (Figure 31). The fitted logistic curves captured the expected sigmoidal shape of categorisation, with the category boundary defined as the stimulus value at which $P(A) = .50$, i.e., 50% probability of a 'Category A' response. The derived boundaries were visualised individually. Additionally, categorisation responses were visualised individually, and response distributions were examined using scatterplots and heatmaps (see Supplemental Materials, Figure S2). Moreover, we conducted a paired-samples t-tests compared category boundaries from Block 1 to Block 2 to assess potential shifts due to prior exposure. Second, we focussed on responses in Block 2 specifically to trials in which the V-H objects lay across the category boundary to examine modality dominance on responses as well as categorisation response bias. In these trials, the visual and haptic shapes belonged to different categories identified by the participant from the congruent trials in Block 1. For example, the V-H conflict information may comprise of the visual shape of object 3 and haptic shape of object 7; with 3 previously categorised as 'A' and 7 categorised as 'B'. Each response was subsequently coded as matching either the visual or haptic category. For each participant and conflict step level, we calculated the proportion of their visual- versus haptic-aligned responses. A binomial generalized linear mixed-effects model (GLMM) tested whether the conflict level predicted visual-dominant responding, with participant as a random intercept.

All analyses were conducted in R using RStudio (R Core Team, 2021). Generalised linear mixed-effects models were fitted with the lme4 package (Bates et al., 2014), and post-hoc comparisons were implemented using the emmeans package (Lenth Russell et al., 2022) and corrected using Bonferroni adjustment. Fixed effects from the final model were converted from log-odds to predicted probabilities using the inverse logit transformation to facilitate interpretation (Muller & MacLehose, 2014). Experiment 7 was preregistered on the Open Science Framework (<https://doi.org/10.17605/OSF.IO/AQ4D5>), and full analysis scripts and data are available at https://osf.io/7wqn4/?view_only=1a9b58fee546486aba1c88b6a8f4f99b.

5.4.2 Result

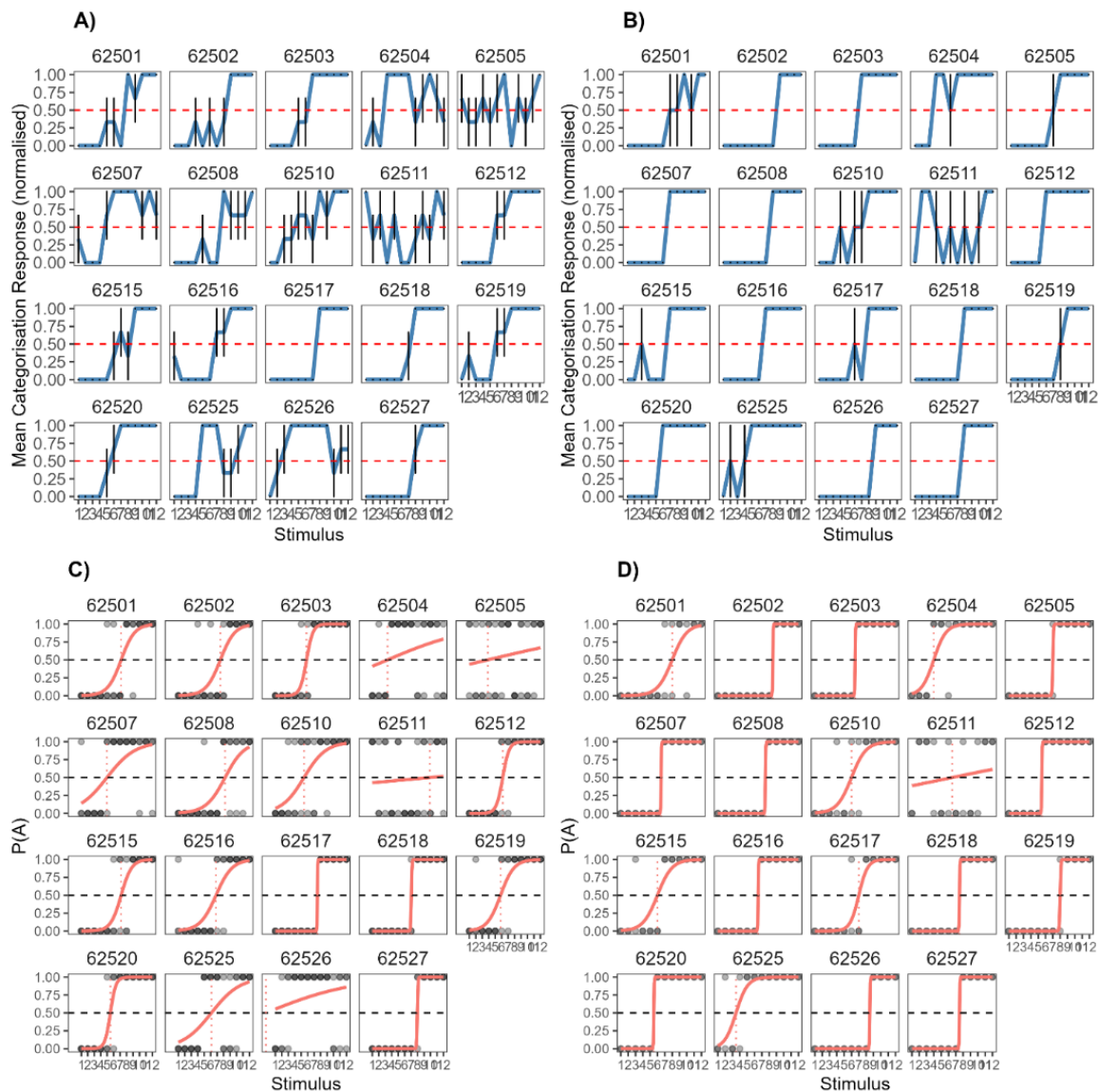
Participants' responses were normalised by coding 'A' responses as 1 and 'B' responses as 0. Participant data was screened for quality including the formation of two distinct categories. Participants who did not show reliable categorisation performance in their responses in Block 1 were excluded from further analyses. Three participants were excluded for this reason, leading to a total sample of 19 participants (6 for object set A; 6 for object set B; 7 for object set C).

Categorisation response to V-H congruent trials

To investigate whether category boundaries remained stable across blocks, participant categorisation responses to congruent trials only were modelled adopting logistic regression, with stimulus number as the predictor and categorisation response (A vs. B) as the outcome. The fitted logistic regression models of categorisation responses confirmed sigmoidal patterns, with steep slopes near category boundaries (see Figure 33). The average position of category boundaries was found between stimuli 7–8 in Block 1 ($M = 7.18$) and 6–7 in Block 2 ($M = 6.47$). We conducted a paired-samples t-test on categories boundaries across blocks, which indicated no overall shift between blocks, $t(16) = -1.56$, $p = .14$, 95% CI $[-1.66, 0.25]$.

Figure 33

Plots showing the raw data and the logistic regression fits on each participant's categorisation responses across Blocks in Experiment 7.



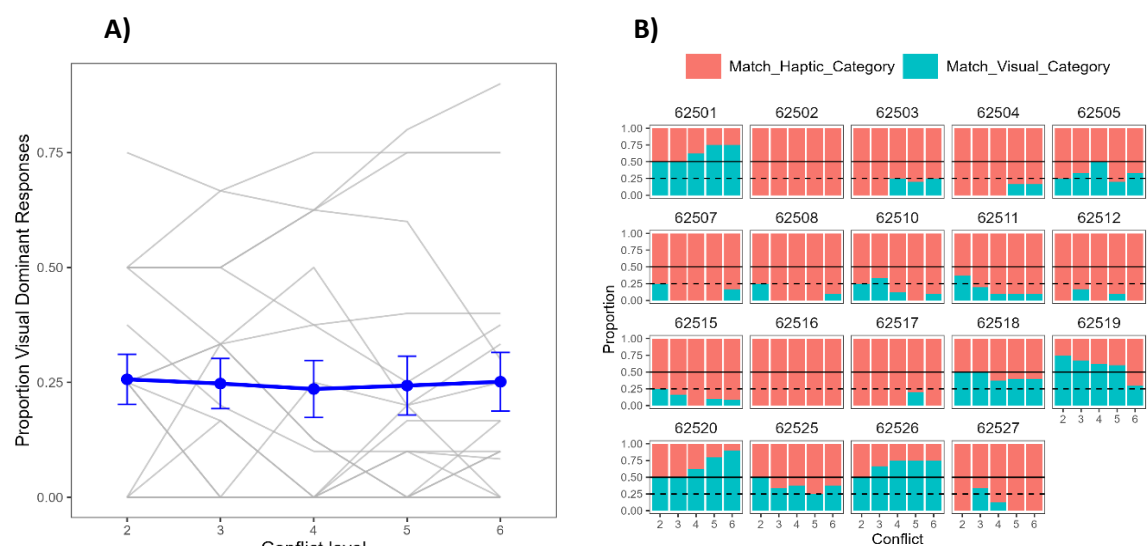
Note: On the top row, the plots show the raw participants' categorisation responses across for Block 1 (panel A) and Block 2 (panel B) in blue lines and with ± 1 SE error bars in dark lines. On the bottom row, the plots show the logistic regression fits on participants' categorisation responses across for Block 1 (panel C) and Block 2 (panel D). Each plot depicts the observed responses (dark and light grey points), fitted curves (red lines), and estimated category boundaries (dotted lines at $P(A) = 0.50$) for each participant.

Categorisation response to V-H incongruent trials

We used a generalised linear mixed-effects model to assess whether the level of shape conflict affected a modality dominance on trials in which the object presented in one modality was categorised differently from the object presented in the other modality (i.e. cross category-boundary trials) only. We conducted a comparison between models with and without the conflict condition. A likelihood ratio test indicated that conflict level did not significantly predict visual-aligned responses on cross-category boundary trials, $\chi^2(4) = 0.44$, $p = .98$. Overall, participants responded in alignment with the visual object on 24.2% of cross-category boundary trials. The mean proportion of visual-dominant responses was stable across conflict levels, although there was substantial variability across participants' categorisation responses (see Figure 34A). The individual-level plots illustrate that some participants were consistently relying on visual versus haptic cues, and some showing mixed patterns (Figure 34B).

Figure 34

The plot shows the category responses to the trials in which the V-H conflict information crossed the category boundary in Experiment 7.



Note: Panel A shows the visual-dominant responses across conflict levels (blue line), with ± 1 SE error bars and individual participant lines (grey). Panel B shows the proportion participant's responses which were aligned with the haptic (red) or visual (blue) object information for each of

the V-H conflict levels (2–6). The solid horizontal lines represent no modality bias (.50) performance in which lower performance represents haptic dominance. For example, and dashed lines at 0.25 represent 25% visual and 75% haptic dominance.

Individual differences investigations

In addition to the preregistered analyses, we conducted separate exploratory analyses to examine individual differences in cue strategy as a function of conflict level (e.g., trial-by-trial reliance on vision vs. haptics as a function of conflict level), and to explore the effect of within- versus between-category boundaries on modality switching. To further explore these individual differences, we classified each participant's response pattern based on their reliance on haptic versus visual information across conflict levels. For each participant, the proportion of responses matching the haptic and visual categories were calculated separately for low (2–3) and high (5–6) conflict trials. Based on these proportions, participants were classified into six strategy groups: haptic-dominant or visual dominant responses which were stable across all conflict levels; a Haptic to Visual switch (haptic dominance at low conflict, visual dominance at high conflict), or Visual to Haptic switch (visual dominance at low conflict, haptic dominance at high conflict), no pattern of sensory dominance. Descriptive statistics indicated that the responses of 12 participants (68.4%) could be classified as Haptic dominant, and 7 participants did not show any sensory dominance. No participant consistently switched between modalities or consistently used visual shape (i.e. visual dominance), for their responses, as reported in Table 8 (for further information about individual strategy refer to Supplementary Materials - Figure S21).

Table 8

Average proportions responses by strategy clusters in Experiment 7.

Strategy	Haptic (M $\pm$ SD)	Visual (M $\pm$ SD)	n
No Clear Dominance	0.49 $\pm$ 0.07	0.53 $\pm$ 0.07	7
Stable Haptic	0.92 $\pm$ 0.06	0.08 $\pm$ 0.06	12

We analysed participants' responses across level of conflict to examine whether there was a reliance on haptic versus visual information. The proportion of responses which matched the category to the haptic information, visual information, or neither modality across each conflict level is shown in Table 9, and suggests consistent dominance of haptic information on category responses across all levels of conflict.

Table 9

Proportion of responses by strategy clusters across conflict levels in Experiment 7.

Conflict Level	RTs (M)	Haptic	Visual	No-Match
Congruent	1683	0.875	0	0.125
Low conflict	1698	0.834	0.069	0.107
Moderate conflict	1692	0.809	0.118	0.072
High conflict	1693	0.765	0.176	0.059

Reaction times

Participant data was screened for time taken to respond to each trial (with outliers defined as ± 3 SD away from the group mean or ± 4 SD away from the individual mean, consistent with previous reports e.g. Williams et al., 2022). This led to a 5.4% of data removed from the RT

data set. A repeated-measures ANOVA on RTs revealed no significant effect of conflict level, $F(3,54)=0.02$, $p=.997$. Across conditions, the mean time taken to categorise the objects was approximately 1.6 seconds and was similar across the different levels of conflict, as reported in Table 7.

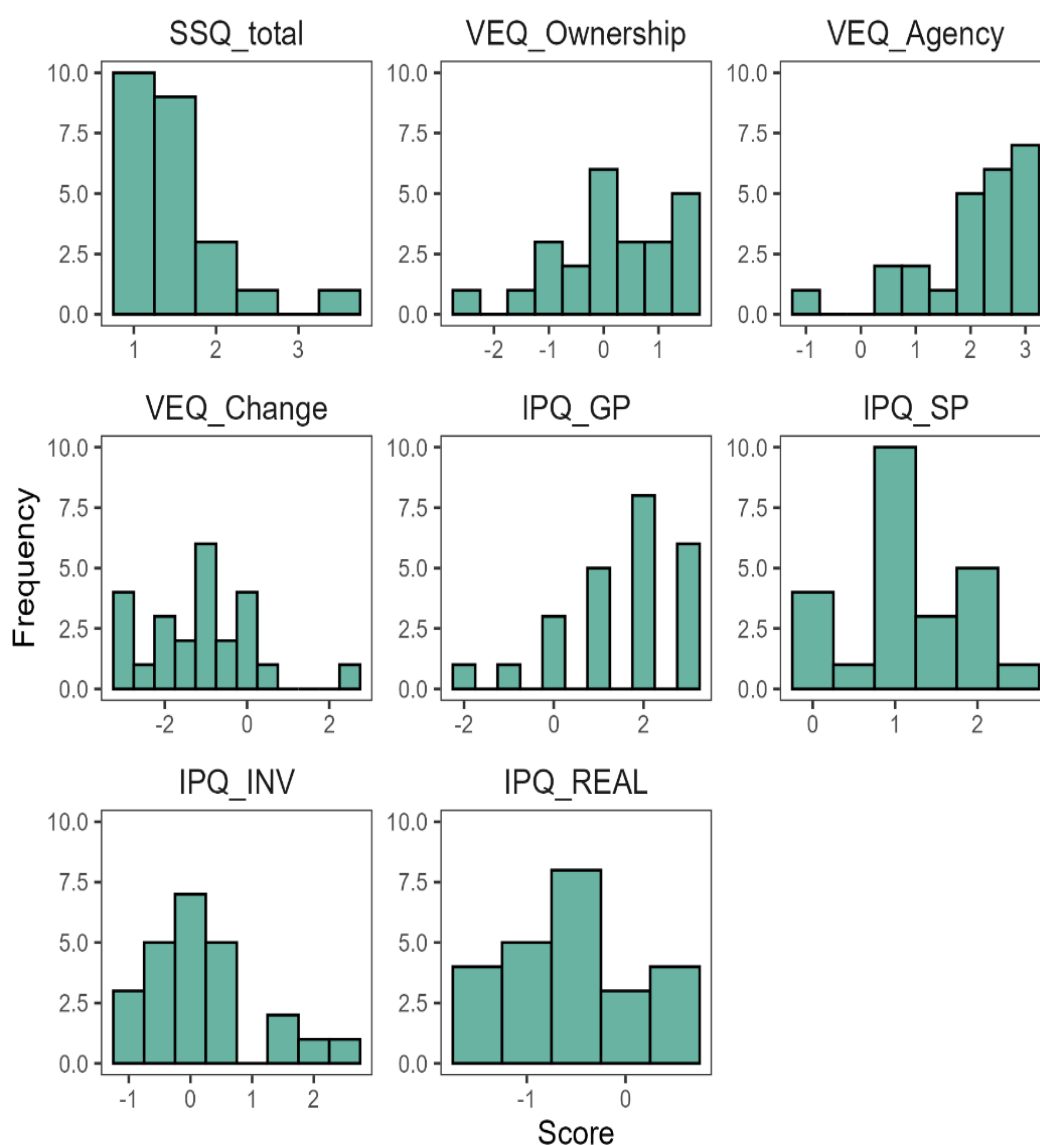
Post-experiment questionnaires

Strategy data. Finally, all participants provided a description, post-experiment of the strategy they adopted while they conducted the task. Participants reported using either both visual and haptic information (26.3%), or haptic information alone (73.7%), and no one reported to have used only visual information. The words 'side', 'sharp', 'edges', 'objects' or 'right' were the most frequently reported, whilst other words (category, hand, left, one, pointy) were less frequently mentioned. Further details can be found under Supplementary Materials (Figure S17).

VRE assessment questionnaires. Participants reported moderate levels of presence in the virtual environment. On the 'Igroup Presence Questionnaire' (IPQ; Schubert et al., 2001; Tran et al., 2024), general presence was moderately high ($M = 1.50$, $SD = 1.35$), spatial presence was also moderate ($M = 1.18$, $SD = 0.70$), but involvement ($M = 0.34$, $SD = 0.90$) and experienced realism ($M = -0.41$, $SD = 0.63$) were low. Simulator sickness, measured by the SSQ (Freiwald et al., 2020; Kennedy et al., 1993), was moderate ($M = 1.47$, $SD = 0.63$). Regarding embodiment, participants reported high agency ($M = 2.15$, $SD = 1.02$), but weak ownership ($M = 0.15$, $SD = 1.05$) and few perceived bodily change ($M = -1.04$, $SD = 1.33$) on the Virtual Embodiment Questionnaire (VEQ; Roth & Latoschik, 2020). Overall, participants' responses ranged from moderate to high, see Figure 35, suggesting individual differences in VR experience.

Figure 35

The distributions of questionnaire scores across participants in Experiment 7.



Note: The histograms show the frequency of responses for each measure, including the Simulator Sickness Questionnaire total score (SSQ_total), Virtual Embodiment Questionnaire subscales (VEQ_Ownership, VEQ_Agency, VEQ_Change), and Igroup Presence Questionnaire subscales (IPQ_GP, IPQ_SP, IPQ_INV, IPQ_REAL).

5.4.3 Discussion

In Experiment 7 we examined how visual and haptic information jointly influence the categorisation of unfamiliar objects, adopting a shape-conflict paradigm within an immersive

virtual reality environment (VR). By parametrically manipulating a visual–haptic shape conflict along a continuum of three-dimensional object shapes, we tested whether categorisation varied with conflict magnitude, whether participants exhibited systematic modality biases, and whether one modality became dominant at higher conflict levels. Our results indicate that participants maintained stable category representations across blocks and predominantly relied on haptic information, even under increasing visual–haptic conflict. This pattern suggests a robust haptic dominance in object categorisation within an immersive VR environment.

On cross-boundary trials, conflict level did not significantly predict visually aligned responses, and most participants showed stable haptic dominance rather than switching modalities or relying primarily on vision. This pattern can be interpreted through the modality appropriateness hypothesis (Welch & Warren, 1980), the modality most suited to the informational demands of the task will dominate processing. While vision is often assumed to be the primary modality for shape perception, haptic exploration can provide complementary and, in some cases, more diagnostic shape information, especially for occluded object features (Newell et al., 2001; Woods & Newell, 2004). The observed haptic dominance may therefore reflect the greater appropriateness of haptic input for resolving shape-based conflicts in this context, extending findings from visuo-haptic object recognition (Klatzky et al., 1985; Lederman & Klatzky, 1987) to categorisation processes. Therefore, the direct, three-dimensional nature of haptic exploration may have provided more detailed shape information than the vision alone, despite the high-fidelity VR environment. Our findings are also consistent with previous evidence showing that haptic information can override visual expectations under conflict or degraded vision (Heller, 1992), especially for material and shape dimensions (Abtahi & Follmer, 2018; List et al., 2014). Haptic dominance has also been observed with blurred visual input, indicating that the perceived reliability of vision can be modulated (Heller, 1992). Thus, although vision is generally more accurate for unambiguous stimuli, touch may provide greater certainty and confidence in

judgements when visual information is unreliable, degraded, noisy or semantically misleading (List et al., 2014; Fairhurst et al., 2018; Abtahi & Follmer, 2018; Girondini et al., 2024).

The nature of the task adopted in Experiment 7 may have biased participants exploration patterns more towards haptics. Categorising unfamiliar three-dimensional shapes requires fine-grained perceptual discrimination, and haptic exploration is good at this (Jones et al., 2005; Lacey & Sathian, 2014). The stability of haptic dominance, even at higher conflict levels, suggests that participants adopted a consistent strategy rather than trial-by-trial strategy for determining category responses. Moreover, the specific stimuli—unfamiliar shapes varying along a morphed shape continuum—may have favoured haptic processing. Haptic object recognition relies on knowledge-driven exploration strategies, where hand movements efficiently extract diagnostic information (Lederman & Klatzky, 1990; 2004). For unfamiliar objects, participants must build representations from bottom-up sensory input rather than stored visual memories (Lacey et al., 2009; Deshpande et al., 2009). Moreover, the parametric manipulation of shape conflict along a continuum may have created conditions in which subtle shape differences were more easily detected haptically than visually (Jones et al., 2005; Gaißert & Wallraven, 2012). Reaction times did not show processing costs, again suggesting efficient conflict resolution, possibly because participants adopted a stable strategy of relying primarily on haptic information. This contrasts with predictions from theoretical perspectives that resolving sensory conflicts should increase cognitive load and slow responses (van Beers et al., 2002; Trommershäuser et al., 2011). However, it is important to note that response times were constrained in our task, in that the participant could respond only following object exploration. It remains possible, therefore, that under more free response time conditions that the effects of cognitive load would be observed.

Unlike traditional laboratory setups with stimuli under highly controlled conditions, our immersive VR environment (VE) provided a more ecologically valid context for studying multisensory integration (Parsons, 2015). However, despite its high visual fidelity, the VE may have

introduced features that effectively reduced the perceived reliability of visual information, leading participants to favour the more direct and less ambiguous haptic input. In Experiment 7, participants reported moderate 'general presence' and 'spatial presence' and a 'sense of control' over their virtual bodies (high agency), but limited 'realism', 'involvement' and 'ownership'. These reports highlight the importance of VR design choices, particularly hand representation, in shaping the embodied experience. The virtual hands adopted in this study were transparent to allow for visual perception of the object shape while haptically exploring the object, but they may have been perceived as unrealistic. The sense of embodiment in VR is important for ownership and affordance judgements (Venkatakrishnan et al., 2023). While full-body avatars can enhance embodiment, research shows that even virtual arms alone, but not hands, can evoke a greater sense of ownership and agency (Slater et al., 2008; 2009). However, the realism of virtual hands strongly influences embodiment: more realistic hands increase embodiment and ownership (Ogawa et al., 2019), whereas transparent hands, although improving object visibility, reduce realism (Zhang et al., 2023). That is, participants may have relied more on haptic than visual object information due to limited realism and embodiment, alongside the belief that touch provides more certain information than vision (Katz, 1989; Heider, 1959; Russell, 1918; Müller, 2013; Fairhurst, et al, 2018). The discrepancy between agency and ownership suggests that participants could control their virtual hands effectively without fully identifying it as their own body. Low realism and involvement further imply that the VR environment was not immersive enough to fully engage participants. Overall, these findings show the importance of enhancing presence, realism and embodiment in VR to support perceptual unity and effective cross-modal integration. Future research should therefore focus on optimising avatar design, particularly hand representation, to strengthen embodiment experiences (Kilteni et al., 2015; Slater et al., 2008; Ogawa et al., 2019).

While this study provides insights into how conflicting shape information across vision and touch is resolved for categorisation, some limitations remain and should be considered. First, it is

known that object familiarity modulates perception (Bulthoff & Newell, 2006; Lacey et al., 2009; Deshpande et al., 2009): familiar objects benefit from prior knowledge that can facilitate cross-modal mapping, while unfamiliar objects rely more on the available sensory information. Our use of unfamiliar objects, while avoiding confounds of prior knowledge, may limit the generalisability of the findings to everyday objects. Future studies should test whether similar patterns of haptic dominance occur with familiar objects, where top-down visual imagery and memory retrieval may alter modality reliance (Lacey et al., 2009). Second, future research could investigate the role of modality specific cues other than shape—such as texture, temperature and compliance in haptics, or colour in vision— in categorisation responses. Third, the fidelity of virtual hand representations can influence the perceived ownership and affordance judgements (Venkatakrishnan et al., 2023). Thus, the use of more realistic hand representations in VR may shift modality reliance and strengthen unity assumptions. The limited realism reported by the participants may constrain the generalisability of our findings to real-world scenarios. Fourth, although the task adopted here, i.e., a categorisation task, reflects an everyday cognitive ability, it may not be the most appropriate to investigate visual–haptic conflict and future studies could adopt a cross-modal matching or discrimination task, which necessitate active integration or comparison of both sensory inputs. Participants would directly compare visual and haptic information about the same object properties, making the conflict more explicit and forcing participants to resolve discrepancies on a trial-by-trial basis (Norman et al., 2004). Finally, individual differences in crossmodal strategies used when there is a conflict in shape should be explored, as preferences for object versus spatial imagery can influence how individuals process multisensory information (Lacey & Sathian, 2014). These individual differences could potentially modulate the degree of haptic dominance observed.

5.5 Conclusion

Overall, our findings demonstrate that the categorisation of unfamiliar objects under visuo-haptic shape conflict remains stable, with no evidence for a shift in category boundaries.

Moreover, categorisation of objects predominantly relied on haptic information, even when visual-haptic conflict increased parametrically and in a virtual reality environment. These findings highlight how conflicts in shape might result in processes that prioritise stable category representations over moment-to-moment perceptual decisions. The robust evidence for haptic dominance observed across all conflict levels shows that touch can play an important role in object perception and categorisation, especially for complex and unfamiliar shapes that require fine-grained discrimination. Moreover, these findings support the need to consider not only sensory information but also task demands and stimulus properties when investigating categorisation following crossmodal conflict in shape. More broadly, our study underlines the stable role of haptic information in shaping object categorisation and the value of VR as a tool to investigate multisensory integration in naturalistic settings.

Chapter 6

General discussion

6 General Discussion

6.1 Introduction

The literature discussed in Chapter 1 supports the integration of several sensory features for both category discrimination and generalisation to novel exemplars (Goldstone & Hendrickson, 2010; Pérez-Gay et al., 2017). Moreover, category specificity is an organisational principle in the brain across visual (Peelen & Downing, 2017), auditory (Feng et al., 2021), and tactile (Lacey et al., 2009) sensory modalities. However, real-world object perception is inherently multisensory (Shams & Seitz, 2008), involving concurrent visual, auditory and tactile inputs; yet despite evidence for multisensory interactions in the brain (Ghazanfar & Schroeder, 2006; Alais et al., 2010), their contributions to object category learning and generalisation remain poorly understood (Newell et al., 2023). To address this knowledge gap, the work presented in this thesis aims to elucidate the role of multisensory perception in the formation of object categories and whether multisensory information leads to the formation of more robust categorisation and efficient generalisation performance. Specifically, we addressed the contribution of visual, auditory and tactile information in multisensory perceptual categories across both familiar and unfamiliar objects.

This chapter summarises and integrates the findings from the preceding experimental chapters (Chapters 2-5) to address whether robust object categorisation and efficient generalisation are influenced by the addition of cues to object shape. These clues included object-related sounds, motion cues, both as tactile and auditory cues, and the redundant shape information through both vision and touch. This discussion further considers the role of object familiarity and top-down influences in shaping the multisensory categorisation processes. Furthermore, the results will be discussed in light of the broader theoretical debate on whether multisensory categories support either early versus late integration, and in turn modality-specific

or amodal object representations. Finally, this chapter concludes by outlining the limitations of the presented work and directions for future research.

6.2 Summary of Main Findings

In Chapter 2, we examined whether prior audio-visual multisensory exposure would modulate the subsequent categorical perception of familiar objects. Although our auditory stimuli were predominantly perceived as semantically matching their corresponding object category (Experiment 1A), we found no multisensory benefit on the category boundary (Experiment 1B). Moreover, participants still categorised objects, particularly for within-category items, as belonging to distinct categories, even in the absence of increased discriminability at the category boundary (i.e., no evidence for categorical perception; Experiment 1B).

In Chapter 3, we investigated whether sound and motion cues influence the perceived shape of unfamiliar objects through inter-object similarity judgements. Principally, object motion is integrated into object shape representation (Experiment 2A), with evidence for separable processing of shape and integral processing of motion (Experiment 3), while auditory cues played a modulatory role under the specific task conditions (Experiment 2B). Overall, findings from Chapter 3 support shape and motion contributing to perceptual similarity, and in turn to the categorisation of objects.

Moreover, in Chapter 4, we investigated whether visual and tactile object motion cues affect the categorisation of unfamiliar object shapes, and whether the formation of multisensory categories supports generalisation to novel exemplars. Results demonstrate that multisensory motion cues support and promote both object category formation as well as generalisation, with effectiveness influenced by cue reliability and task structure. Specifically, performance improved when all three-cues available at test for both all cues fully predictive (Experiment 4) and shape reliability reduced (Experiment 5) for category membership. Task format modulated these benefits: blocked presentation (Experiments 4A and 5B) showed higher accuracy and stronger

generalisation, whereas interleaved presentation (Experiments 4B and 5B) impaired performance and weakened tactile cues contributions.

Finally, in Chapter 5 we explored visuo-haptic conflict resolution during categorisation of unfamiliar objects in virtual reality (VR). Results show that participants retained stable category boundaries and largely relied on haptic shape information, even when visual-haptic conflict increased parametrically. This finding suggests that visuo-haptic categorisation is biased towards haptic input, which may have provided more detailed shape information than the visual input via VR; especially for complex and unfamiliar shapes that require fine-grained discrimination. Finally, the consistent categorisation performance despite sensory conflict implies access to shared, modality-independent representations rather than trial-by-trial cue reweighting.

6.3 Implications of the Findings

6.3.1 The visual object shape, but not its sound, support the categorisation of familiar objects

In Chapter 2, we did not find evidence for sound object cues facilitating category representation of familiar objects. However, participants successfully categorised the stimuli based on visual shape alone, even in the absence of categorical perception. This finding aligns with existing evidence that shape serves as a fundamental organising principle for object categories (Biederman, 1987; Tarr & Bülthoff, 1998; Hayward, 1998), due to its invariant properties across viewing conditions and its diagnostic value for object identity (Edelman, 1998; Goldstone & Barsalou, 1998).

Auditory cues did not lead to more robust categories for familiar objects. The question this raises is whether sounds of familiar objects are inherently irrelevant for object categorisation, or whether was this outcome a consequence of the specific experimental design. Although, there is abundant evidence providing support for object sounds being integrated into object representation (Mulatti et al., 2014) and supporting efficient recognition (Murray et al., 2006;

Lehmann & Murray, 2005; Suied & Viaud-Delmon, 2009), the usefulness of sound for categorisation is highly content- (Bones et al., 2018) and context- dependent (Heald & Nusbaum, 2017). Unlike spatial and geometric information processed by object shape, the sounds that objects make are often indicative of their material properties, the actions being performed on them, or their size, rather than their categorical identity (Bizley & Cohen, 2013). Moreover, while some objects have highly diagnostic sounds (e.g., bell and a telephone), many others do not (e.g., a vase), or they produce sounds that are highly variable and dependent on the interaction. Thus, the semantic object sounds adopted in Chapter 2 may not have been sufficiently diagnostic of the object categories tested. In other words, they might not have been characteristic sounds or represent the individual differences in object category representations that emerge through natural multisensory experience (Murray et al., 2016; Matusz et al., 2017; Liu et al., 2019).

Another aspect to consider is whether sounds enhance object categorisation either at early or late stages of processing. There is evidence for object sounds benefiting object perception at the early processing stages (Williams et al., 2022), as well as at the decisional level (Franzen et al., 2020). The object sounds in Chapter 2 were presented prior to the main task, thus might not have had an effect because they were presented at the learning-exposure stage instead of the categorisation one. Moreover, the absence of an effect of sound on categorisation may reflect the task-irrelevant nature of the auditory stimuli, consistent with previous research demonstrating that sounds are effectively integrated when they are semantically congruent or task-relevant (Chen & Spence, 2010; Bizley et al., 2016). Future studies could investigate potential effects of multiple exemplars of the same object's sounds exposed over a longer period (i.e, in different tones, or sound type). Another future direction could be to investigate whether morphed sounds which match the morphed objects allows the formation of a more robust object representation, which in turn might lead to enhanced categorisation performance.

The failure to replicate categorical perception effects for objects previously presented both in audio-visual and visual-only conditions may reflect ceiling effects in discrimination

performance, as participants likely relied on prior knowledge and top-down influences when categorising familiar objects (Goldstone, 1994; Schyns et al., 1998). This interpretation is consistent with evidence that categorical perception emerges more readily with unfamiliar stimuli where participants must rely on perceptual rather than semantic processing (Goldstone & Hendrickson, 2010; Newell & Bülthoff, 2002). Future research could investigate modification of the paradigm by increasing exposure during learning (leading participants to actually learn the objects through feedback) and reducing the time between learning and test. Alternatively, one could either adopt unfamiliar objects or reduce visual reliability, i.e. making visual information more ambiguous, since previous work demonstrated how auditory information support the disambiguation of visual information (Vroomen & de Gelder, 2000), thus becoming task-relevant at learning.

6.3.2 The role of visual object shape and audio-visual object motion cues on inter-object perceived similarity of unfamiliar objects

Findings from Chapter 3 demonstrated that visual motion cues systematically affected the perceived similarity and classification of unfamiliar objects, while auditory motion cues did not. These findings extend research showing that motion information is integrated into object identity representations (Vuong & Tarr, 2004; Wallraven & Bülthoff, 2009) and support the view that dynamic properties become part of the stored object representation in memory (Newell, et al., 2004; Setti & Newell, 2010; Chung et al., 2023). Similarity ratings were adopted as an indirect measure of categorisation processes, grounded in the theoretical assumption that discrimination among stimuli guides and reflects category formation, such that perceived similarity reveals the latent discriminative structure (Shepard, 1987; Goldstone, 1994). The analysis of similarity ratings, therefore, allows inference of the discriminative basis of categorisation, linking perceptual judgments to category formation processes (Ashby & Lee, 1991).

The effectiveness of visual motion found in Chapter 3 reflects its ability to provide dynamic depth information that reveal three-dimensional object structure (Braunstein et al., 1990; Norman & Todd, 1993; Newell et al., 2004). Indeed, object motion provides support for resolving structural ambiguity and facilitating identification by revealing additional object properties (Norman et al., 2004; Newell et al., 2004). Moreover, the effect of visual motion appears to be content specific: while objects moving in the same way appear more similar, objects moving in a different way become less similar. Additionally, NMDS space reconstructions and clustering suggest that visual motion led to overall more coherent perceptual grouping (Grossman & Blake, 1999), in a content-specific manner. Finally, the asymmetrical interference between shape and visual motion dimensions observed suggests that motion serves as an integral feature in the representational space, consistent with evidence for dimensional integration/modulation in object perception (Folstein et al., 2013).

In Chapter 3, auditory motion cues did not affect similarity rating for unfamiliar object, aligning with the view that object sound cues do not play a functional role in the initial recognition process unless it is semantically relevant and part of a stored, pre-existing concept (Chen & Spence, 2010; Mulatti et al., 2014). Although the human brain is capable of categorising objects based on sound alone (Murray et al., 2006; Iamshchinina et al., 2022), this is both context- and content (i.e. informativeness relevant for categorisation)-dependent, as discussed previously. Thus, while object shape and visual motion are often stable and central properties to infer object's identity (Johannsen 1973; Newell et al., 2004), motion sounds may act as transient or incidental cues. Thus, both Chapter 2 and 3 support a cue-reliability hierarchy (Ernst & Banks, 2002) in the brain to prioritise inputs that are diagnostic and stable features of an object's identity (Schyns et al., 1998).

6.3.3 *Beyond Shape: the role of visuo-tactile motion cues in categorising multisensory unfamiliar objects*

In Chapter 4, we found compelling evidence that both visual and tactile motion cues support more efficient categorisation and generalisation of unfamiliar objects. Alongside Chapter 3, this finding extends research on object motion being both sufficient (Johannsen, 1973) and integrated in object identity (Newell et al., 2004) for efficient categorisation. Moreover, these findings further extend previous work on visuo-haptic object recognition (Lacey & Sathian, 2014; 2020; Snow et al., 2014), suggesting complementarity between visuo-haptic modalities (Ernst & Banks, 2002; Helbig & Ernst, 2007; Woods & Newell, 2004) and that motion cues can be effectively integrated across modalities for efficient categorisation. Specifically, the efficacy of haptic motion cues aligns with the view that touch provides rich information about object dynamics (Lederman & Klatzky, 2009; Klatzky & Lederman, 2011). Furthermore, the convergence of dynamic information across modalities supports the view that motion provides a robust information for multisensory object categorisation (Lacey & Sathian, 2014).

The presentation format effects observed in Chapter 4 also align with research on contextual modulation of multisensory integration (Rohe & Noppeney, 2015; Ten Oever et al., 2016). Under blocked conditions, motion cues may function as contextual supports in memory for categorising object shape (Duarte et al., 2023, 2025), while interleaved presentation may cause greater reliance on individual cues or impair retrieval of specific cue-category associations (Carvalho & Goldstone, 2014; 2015). We found this pattern to be consistent across categorisation accuracy, generalisation performance, category boundary discrimination, and category representations. Considering category representations, category reconstructions varied by presentation format: blocked presentations produced clearer category distinctions than interleaved ones. Specifically, during interleaved presentation, participants may struggle to extract and match individual object cues with those learned, leading to greater reliance on abstract category representations. In other words, participants might struggle to extract individual cues when conditions are interleaved, due

to increased unpredictability. While such abstraction could be less precise, it enables broader grouping of stimuli into either category. This suggests that under increased task uncertainty, the brain may prioritise abstract representations of multisensory categories over specific cue extraction. This interpretation is further supported by boundary analyses: removing cues relative to the three-cue learning condition led to shallower slopes. This indicates that participants' category decisions efficiency relied on the presence of all learned multisensory cues. Consistent with evidence of late integration (Williams, et al., 2022; Li et al., 2024), this would suggest that multisensory learning produces robust but less dissociable category representations. These representations, while supporting strong generalisation, become vulnerable when individual cues are selectively removed.

Finally, the paradigm adopted in Chapter 4 is based on learning multisensory (three-cues) categories, followed by categorisation under cue removal at test. Overall, the results indicate that multisensory learning creates robust category representations that are difficult to deconstruct into individual cue components, as shown by a significant decreased categorisation efficiency. Moreover, when participants learned the categories through shape alone, the additional cues at test did not enhance performance, likely due to their redundancy, as shape information was sufficient. This pattern supports late-integration models where multisensory features become bound into coherent category representations that resist segregation (Li et al., 2024). The superior performance observed for three-cues at test may also reflect a familiarity effect, with the same cues being presented at both learning and test (Duarte et al., 2023). Importantly, multisensory learning resulted in enhanced generalisation irrespective of varying cues diagnostic probability for category membership (Experiments 4 and 5), suggesting that robust category formation is not solely attributable to familiarity or repetition, but rather to the integrative effect of multisensory learning.

6.3.4 *Haptic dominance in shape conflict for categorisation of unfamiliar objects*

Chapter 5 revealed haptic dominance in shape-based categorisation when both visual and haptic information were available, despite predictions from optimal cue integration models favouring visual dominance (Ernst & Banks, 2002; Landy et al., 1995, Snow et al., 2014). This finding challenges the assumption that sensory weighting follows firm reliability principles and suggests dynamic integration where contextual factors can override optimal integration strategies.

The haptic dominance observed in Chapter 5 may be interpreted through top-down or bottom-up processes. This may reflect context-dependent bias, where the virtual reality environment increased reliance on haptic shape information due to reduced visual credibility (Miller & Sheinberg, 2022). In the VR environment, visual information may lack ecological validity, whereas haptic feedback provides more direct spatial cues (Slater et al., 2008; Riva et al., 2019). It is also possible that this pattern reflects a top-down, attention-driven bias, whereby visual cues originating from the virtual environment are less attended to or less informative, further enhancing the relative contribution of haptic input. Alternatively, haptic dominance may also reflect access to complementary shape information unavailable through vision alone, such as hidden surfaces of objects (Newell et al., 2001; Woods & Newell, 2004), extending previous findings on visuo-haptic object recognition to categorisation tasks. Finally, haptic dominance might be partially explained by individual differences in learning strategies, as has been previously reported in the acquisition of novel categories (Maddox et al., 2006; Smith et al., 2014). This suggests that individuals differ in the extent to which they depend on specific modalities during and approach learning. Such differences have also been reported in modality preference or effectiveness, with some individuals showing stronger performance in one modality compared to others (Roark, et al., 2021). These individual differences may therefore contribute to variability in categorisation processes and in the formation of multisensory object categories.

Our findings support a nuanced view of multisensory integration, where the brain's response to visual–haptic conflict depends on the assumed unity of sensory signals, their relative reliability, and specific task demands (Ernst & Banks, 2002; Helbig & Ernst, 2007). In tasks requiring fine-grained geometric judgements, such as the shape categorisation employed in Chapter 5, haptic information may be inherently more reliable than visual input, leading to systematic haptic dominance regardless of conflict magnitude. The stability of categorisation boundaries across experimental blocks suggests that participants maintained consistent category representations despite conflicting visual–haptic information.

6.3.5 Theoretical implications: toward a continuum approach for multisensory object representations

In multisensory research, two longstanding debates have shaped the nature of multisensory interactions: the modal versus amodal representations debate (Barsalou et al., 2003; Ghazanfar & Schoreder, 2006; Kaup et al., 2023; Mahon & Caramazza, 2008; Mahon, 2015; Miller & Sheinberg, 2022), and the early versus late integration debate (Meredith & Stein, 1983; Calvert et al., 2004; Foxe & Schroeder 2005; Beauchamp et al., 2008). While traditionally considered in isolation, these debates converge on a common theoretical question, such as when and in which form sensory signals are integrated into coherent perceptual and conceptual representations. The findings discussed in the previous chapters challenge this binary distinction and provide evidence supporting a continuum-based framework for understanding multisensory object categorisation (Patterson et al., 2007; Lambon Ralph et al., 2017). Additionally, our findings build on the view of flexible, adaptive mechanisms in multisensory perception (Noppeney, 2021; Rohe & Noppeney, 2015).

The longstanding debate on the nature of object representations concerns whether they are stored as modality-specific and as such are grounded in sensory and motor systems (modal; Barsalou, 1999; 2008; Pulvermuller, 2013; Mahon, 2015) or are they abstract, modality-

independent representations (amodal; Fodor, 1983; Pylyshyn, 1984). On one hand, evidence for modal-format arises from neuroimaging studies finding modality-specific activation for thinking of concepts (Martin, 2007; Kiefer & Pulvermüller, 2012; Miller & Sheinberg, 2022) and behavioural priming studies showing that exposure to one modality speeds recognition in the other (Chen & Spence, 2011; Brunel et al., 2013). On the other hand, evidence for amodal-format arises from cross-modal transfer categorisation studies (Yildirim & Jacobs, 2013; 2015; Wallraven et al., 2014), and neurobiological evidence for modality-independent, shared shape representations in the LOC (Amedi et al., 2001; 2002) and IPS (Lacey et al., 2009). This suggests that the LOC's function might be "object-based", and processing geometric shape information regardless of how it is perceived (Masson et al., 2016), which aligns with the view of the brain organised in task-based functions with multisensory inputs (Lacey & Santhia, 2014). The modern framework integrates these two extremes, suggesting a hybrid hub-and-spoke architecture (Patterson et al., 2007; Lambon Ralph et al., 2017), where modality-specific "spokes" process sensory information in specialised systems, which connect to an amodal "hub" (anterior temporal lobe, ATL) supporting cross-modal generalisation, semantic integration and conceptual abstraction. This hybrid view reconciles cross-modal behavioural transfer with modality-specific activations, arguing for early representation being modality specific, and higher-level representations be amodal. Evidence for the hub-and-spoke model arises from converging findings across neuropsychological (Lambon Ralph et al., 2009), neuroimaging (Chiou & Lambon Ralph, 2019), and computational modelling (Lambon Ralph & Patterson, 2008) research.

Although not explicitly investigating the nature of object representations, our findings contribute to this debate by providing evidence for hybrid representations operating at multiple levels of abstraction, consistent with the hub-and-spoke model (Patterson et al., 2007; Lambon Ralph et al., 2017). The integration of shape and motion cues (Chapters 3 and 4) for efficient categorisation aligns with the view that properties with high crossmodal redundancy are represented in an integrated or supra-modal manner (Spence & Di Stefano, 2024). Similarly,

results from Chapter 4 support amodal integration, as participants struggled to dissociate learned multisensory categories into separated sensory components. This pattern suggests that multisensory category learning creates integrated representations stored in modality-independent formats. However, the failure of sounds to support efficient categorisation (Chapters 2 and 3) aligns with the view that non-diagnostic cues might remain modality-specific (Roark, 2024) and are activated through an abstract-conceptual link (Mulatti et al., 2014). Likewise, the context-dependent sensory dominance pattern (Chapter 5) suggests the role of modality-specific processing in categorisation performance, with dynamic weighting mechanisms adjusting the relative contribution of the ‘most credible’ sensory modality for the task (Noppeney, 2021). Finally, while one possibility is that memory stores holistic object representations (Martin, 2007; Di Carlo et al., 2012), our findings appear to partially align with accounts proposing that memory primarily encodes features, which may be integrated across modalities to support multisensory object categorisation (Yzenberg & Behrmann, 2022).

Overall, our findings are consistent with contemporary views suggesting a representations continuum for multisensory object categories, ranging from modality-specific sensory information to more abstract, amodal concepts (Michel, 2021; Kaup et al., 2024). It is likely that both modal and amodal representations coexist and are used flexibly depending on the task and the context. That is, multisensory categories are a dynamic system where the degree of integration depends on the specific cue and its diagnostic value. Our findings suggest that reliable, redundant cues like shape and motion are integrated into a supra-modal format, while other, less diagnostic cues may remain tied to their originating modality to support efficient, flexible categorisation and generalisation performance.

Regarding multisensory integration timing, neurophysiology and neuroimaging research shows how early multisensory integration (<100 ms) in the primary sensory cortices provides benefit to detection and localisation (Foxy & Schroeder, 2005; Van der Burg et al., 2011). By

contrast, later interactions (> 250 ms) in multisensory convergence zones such as STS, IPS and prefrontal and temporal cortices, support binding (Calvert et al., 2001; Beauchamp et al., 2004), attentional modulation (Talsma et al., 2010; Donohue & Woldorff, 2011), semantic recognition (Beauchamp, 2005) and categorisation (Viganò et al., 2021; Doehrmann & Naumer, 2008). The temporal and anatomical dissociations between early and late interactions (Talsma et al., 2010) elucidate their complementary functions, as both operate within a hierarchical and distributed system where integration occurs at multiple stages of processing (Stein & Stanford, 2008; Ghazanfar & Schroeder, 2006). The distributed nature of these interactions allows modality-specific information to be rapidly combined for perception while being available for higher-order, amodal or hub-mediated conceptualisation (Gau et al., 2020). Although not specifically investigated in this work, the findings presented could support the continuous integration view serving different functions, determined by task demands and contextual variations, which might lead different paradigms to yield different results. For instance, while spatial features (shape, motion) suggest early integrations, the semantic features (contextual associations) might suggest late integration to support efficient categorisation.

6.3.5.1 Task demand specificity and information type

The variable effectiveness of different sensory cues (i.e., spatial versus temporal information) might depend on both task demands and the specificity of modality information. While vision and haptics excel at processing spatial information (O'Regan & Noë, 2001), audition specialises in temporal pattern decoding (Griffiths & Warren, 2004; Zatorre et al., 2002). Thus, it would be reductive to attribute the absence of auditory cues benefit to a general inability of auditory information contributing to multisensory categorisation. Alternatively, it might reflect a discrepancy between task demands (spatial categorisation) and auditory processing strengths (Kubovy & Van Valkenburg, 2001; Spence, 2007), supporting the modality appropriateness account (Welch & Warren, 1980). This interpretation aligns with evidence of multisensory integration being constrained by the informativeness and consistency of individual sensory

features (Shams & Beierholm, 2010; Kayser & Shams, 2015). Results from Chapter 4 suggest that integration processes can be flexibly reconfigured based on task demands and learning context. After learning multisensory categories, participants likely developed integrated representations that were difficult to dissociate, suggesting that the brain can adaptively organise integration processes based on i) learning (multisensory vs. unisensory categorisation), ii) presentation format, iii) cue availability and iv) generalisation demands (learned vs. novel stimuli). This flexibility supports predictive coding models of multisensory integration (Friston, 2010; Talsma, 2015), where the brain continuously updates integration strategies based on environmental statistics and task demands.

Overall, while visual shape cues benefited from simultaneous access to multiple spatial features, auditory objects relied primarily on temporal dynamics that did not provide sufficient diagnostic information to support categorisation. This highlights modality-specific qualities: vision allows rapid global processing of spatial features, whereas auditory cues may require more naturalistic contexts to become effective. Importantly, our results extend the view that multisensory information leads to more robust object representations (Ernst & Bühlhoff, 2004) by demonstrating that multisensory information also supports the formation of more robust categories of novel objects. As novel objects are less influenced by prior semantic knowledge or top-down expectations, these findings provide strong evidence that the robustness of categorisation originates from the integration of multiple bottom-up sensory cues. In line with hybrid hub-and-spoke models (Lambon Ralph et al., 2017; Kaup et al., 2024) and graded representational frameworks (Michel, 2021), we propose that reliable, redundant cues such as shape and motion are integrated into a supramodal representational format, whereas non-diagnostic cues remain modality-specific. Precision and weighting mechanisms might guide whether multisensory signals remain bound to their sensory origin or abstracted into modality-invariant categories, depending on task demands and cue properties (e.g., spatial vs. temporal). For example, when categorisation depends on shape, participants prioritise spatially diagnostic

visual or haptic information, whereas auditory cues contribute less. Overall, these findings suggest that multisensory information contributes to the formation of robust object categories of unfamiliar objects.

6.4 Novel Contribution of this Thesis

This thesis introduces both methodological and theoretical innovations that advance research on multisensory object perception and categorisation. Regarding theoretical aspects, this thesis extends previous work on multisensory object recognition toward the study of multisensory categorisation processes. Building on findings that multisensory information supports more robust object representations (Ernst & Bühlhoff, 2004), we show that multisensory input enhances the formation of categories for unfamiliar objects. Moreover, unlike most research in the categorisation domain, which often relies on low-level stimuli (i.e. Gabor patches or pure tones), the work presented here employs ecologically valid, high-level familiar and unfamiliar objects, capturing the complexity of real-world objects. The approach adopted in this thesis allows for the assessment of how object features are integrated in a multisensory manner to support category formation.

Methodologically, this thesis extends previous work on online multisensory experimentation (Weydmann et al., 2023; O'Dowd et al., 2025; Hirst et al., 2025) by demonstrating the feasibility of conducting controlled multisensory experiments remotely through participants' own devices, including smartphones and desktops. Particularly, experiments reported in Chapters 2–4 were conducted through online recruitment, delivering complex visual, auditory, and tactile (vibration) stimuli with precision (Anwyl-Irvine et al., 2021; Hirst et al., 2025). Although online environments are inherently noisier and less controllable than laboratories (Bridge et al., 2020), strong multisensory effects nonetheless emerged (see Chapter 4), this suggests that the effects we found are likely to be at least as strong, if not stronger, under controlled laboratory conditions. While psychophysical and neuroimaging studies often require

laboratory comparisons to ensure stimulus precision, our investigation of multisensory category processing demonstrates that reliable effects can be obtained in more ecologically valid, real-world environments. These methodological novelties provides the link to bridge the gap between strictly controlled psychophysical paradigms and naturalistic sensory experiences, enhancing the generalisability of findings to an applied setting. Online testing thus offers a scalable, reproducible, and accessible approach without compromising validity, enabling other researchers to replicate and extend these findings.

Furthermore, this thesis provides a battery of openly accessible online experiments and stimuli for future research in multisensory perception and categorisation. Specifically, we developed a novel stimulus set comprising three-dimensional, parametrically morphed objects, which can also be 3D-printed to allow for haptic experimentation (see Chapter 5). This dataset has been made publicly available on the OSF repository (https://osf.io/7wqn4/?view_only=1a9b58fee546486aba1c88b6a8f4f99b) and supports future investigations of these complex familiar and unfamiliar stimuli across visuo-haptic modalities. By openly sharing these materials, the present work facilitates replication and the development of standardised tools for multisensory research.

6.5 Limitations of the Current Work and Future Directions

The present thesis provides novel insights into multisensory object categorisation; however, several limitations should be considered. First, the experiments primarily focused on unfamiliar objects (Chapters 3–5), with limited investigation of familiar stimuli (Chapter 2). We mainly focused on unfamiliar objects to investigate the role of bottom-up processes while minimising the influence of prior knowledge, which has been shown to modulate multisensory processing and categorisation (Talsma et al., 2010; Liu et al., 2011). Future research could examine whether shape- and motion-based multisensory cues similarly influence categorisation of familiar objects, clarifying interactions between prior experience and multisensory perception in

categorisation processes. Second, auditory stimuli varied across experiments: Chapter 2 adopted semantically related sounds, while Chapter 3 employed non-semantic sounds. Future investigations could systematically pair semantic sounds with unfamiliar objects and comparing these with non-semantic sounds to investigate their relative contribution to category learning. Additionally, incorporating other object cues (i.e. material properties) or using multiple sound variants (i.e., morphed or statistically varied exemplars) could help clarify the conditions under which auditory cues contribute effectively to object categorisation. Third, the category learning processes employed in this thesis has been restricted to rule-based (RB) category paradigms, which while informative, may not generalise to information-integration (II) learning strategies that engage distinct neural mechanisms. Furthermore, previous work has reported individual differences in successfully acquiring II category structures (Maddox et al., 2006; Smith et al., 2014; Roark & Chandrasekaran, 2023). Given these challenges, the present work focused on RB learning to ensure reliable acquisition of category rules. Thus, the results presented here are specific to unfamiliar objects learned with RB category learning and are potentially not generalisable to other types of category-learning paradigms. Future research should examine II paradigms with multisensory objects to test whether observed effects generalise to more complex category structures.

Fourth, the present work primarily focused on bottom-up perceptual processes influences on multisensory object categorisation, without top-down manipulations such as prior knowledge, expectation, or attentional modulation. Future studies should integrate top-down factors to assess cognitive-perceptual interactions in multisensory categorisation. Similarly, we exclusively relied on behavioural measures. While behavioural data provides valid and informative insights, combining behavioural paradigms with neuroimaging, EEG/ERP and connectivity analyses could provide a stronger neural basis to elucidate the temporal dynamics of multisensory integration in categorisation processes. Particularly, this combination of methodologies could clarify shared versus modality-specific neural representations and the plasticity which supports flexible sensory

weighting during categorisation processes. Finally, although sensorimotor experiences, motor engagement, and action planning can shape multisensory object categorisation (Iachini et al., 2008; Yang et al., 2017; Camus et al., 2018), the role of embodied cognition was not explored here. Future research incorporating embodied tasks may further clarify interactions between perception, action, and multisensory category learning.

6.6 Conclusion

This thesis advances our understanding of how multisensory information contributes to object categorisation and generalisation processes. The findings demonstrate that sensory cues differ in their effectiveness: while visual shape cues benefited from simultaneous access to multiple spatial features (i.e., visual and haptic inputs), auditory cues, which rely primarily on temporal dynamics, did not provide sufficiently diagnostic information to support categorisation. Our findings are consistent with contemporary perspectives suggesting a representational continuum for multisensory object categories, ranging from modality-specific sensory information to more abstract, amodal concepts. By extending previous work on multisensory object recognition, this thesis shows that multisensory input enhances the formation of categories for unfamiliar objects. Furthermore, through the adoption of online and ecologically valid methodologies, and by providing a set of openly accessible experiments and stimuli, this research facilitates replication and the development of standardised tools for the field. Looking ahead, future research should move beyond the bottom-up perceptual focus of the present work by incorporating top-down factors, and by combining behavioural paradigms with neuroimaging approaches to provide a stronger neural basis for understanding the temporal dynamics, representational overlap, and plasticity underlying flexible multisensory categorisation.

7 References

- Abel, R. (2023). Interleaving effects in blindfolded perceptual learning across various sensory modalities. *Cognitive Science*, 47(4), e13270. <https://doi.org/10.1111/cogs.13270>
- Abel, R., Brunmair, M., & Weissgerber, S. C. (2021). Change one category at a time: Sequence effects beyond interleaving and blocking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(7), 1083 – 1105. <https://doi.org/10.1037/xlm0001003>
- Abtahi, P., & Follmer, S. (2018). Visuo-haptic illusions for improving the perceived performance of shape displays. In *proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-13). <https://doi.org/10.1145/3173574.3173724>
- Adam, R., & Noppeney, U. (2009). Visual object categorization with conflicting auditory information. *NeuroImage*, 47(S1), S39-S41. [https://doi.org/10.1016/S1053-8119\(09\)70022-9](https://doi.org/10.1016/S1053-8119(09)70022-9)
- Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3), 257-262. <https://doi.org/10.1016/j.cub.2004.01.029>
- Alais, D., Newell, F. N., & Mamassian, P. (2010). Multisensory processing in review: from physiology to behaviour. *Seeing and Perceiving*, 23(1), 3-38. <https://doi.org/10.1163/187847510x488603>
- Amedi A, Jacobson G, Hendler T, Malach R, Zohary E. (2002). Convergence of visual and tactile shape processing in the human lateral occipital complex. *Cereb Cortex*, 12, 1202–1212. <https://doi.org/10.1093/cercor/12.11.1202>
- Amedi A, Malach R, Hendler T, Peled S, Zohary E. (2001). Visuo-haptic object related activation in the ventral visual pathway. *Nat Neurosci*, 4, 324–330. <https://doi.org/10.1038/85201>
- Amedi, A., Jacobson, G., Hendler, T., Malach, R., & Zohary, E. (2002). Convergence of visual and tactile shape processing in the human lateral occipital complex. *Cerebral Cortex*, 12(11), 1202-1212. <https://doi.org/10.1093/cercor/12.11.1202>
- Amemiya, T., Beck, B., Walsh, V., Gomi, H., & Haggard, P. (2017). Visual area V5/hMT+ contributes to perception of tactile motion direction: a TMS study. *Scientific Reports*, 7(1), 40937. <https://doi.org/10.1038/srep40937>
- Angelaki, D. E., Gu, Y., & DeAngelis, G. C. (2009). Multisensory integration: psychophysics, neurophysiology, and computation. *Current Opinion in Neurobiology*, 19(4), 452-458. <https://doi.org/10.1016/j.conb.2009.06.008>

- Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior research methods*, 53(4), 1407-1425. <https://doi.org/10.3758/s13428-020-01501-5>
- Ashby, F. G., & Lee, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General*, 120(2), 150–172. <https://doi.org/10.1037/0096-3445.120.2.150>
- Ashby, F. G., & Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology*, 38(4), 423-466. <https://doi.org/10.1006/jmps.1994.1032>
- Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol.*, 56(1), 149-178. <https://doi.org/10.1146/annurev.psych.56.091103.070217>
- Ashby, F. G., & Spiering, B. J. (2004). The neurobiology of category learning. *Behavioral and Cognitive Neuroscience Reviews*, 3(2), 101-113. <https://doi.org/10.1177/1534582304270782>
- Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, 93(2), 154–179. <https://doi.org/10.1037/0033-295X.93.2.154>
- Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In *Handbook of Categorization in Cognitive Science*, 157-188. <https://doi.org/10.1016/B978-0-08-101107-2.00007-5>
- Ashby, F. G., Alfonso-Reese, L. A., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, 105(3), 442-481. <https://doi.org/10.1037/0033-95X.105.3.442>
- Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K., & Bizley, J. K. (2018). Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding. *Neuron*, 97(3), 640-655. <https://doi.org/10.1016/j.neuron.2017.12.034>
- Atkin, C., Stacey, J. E., Roberts, K. L., Allen, H. A., Henshaw, H., & Badham, S. P. (2023). The effect of unisensory and multisensory information on lexical decision and free recall in young and older adults. *Scientific Reports*, 13(1), 16575. <https://doi.org/10.1038/s41598-023-41791-1>
- Atkins, J. E., Fiser, J., & Jacobs, R. A. (2001). Experience-dependent visual cue integration based on consistencies between visual and haptic percepts. *Vision Research*, 41(4), 449-461. [https://doi.org/10.1016/S0042-6989\(00\)00254-6](https://doi.org/10.1016/S0042-6989(00)00254-6)
- Audacity Team. (2023). *Audacity (Version 3.4.2) [Computer software]*. <https://www.audacityteam.org/>

- Ayzenberg, V., & Behrmann, M. (2022). Does the brain's ventral visual pathway compute object shape?. *Trends in Cognitive Sciences*, 26(12), 1119–1132. <https://doi.org/10.1016/j.tics.2022.09.019>
- Ayzenberg, V., Chen, Y., Yousif, S. R., & Lourenco, S. F. (2019). Skeletal representations of shape in human vision: Evidence for a pruned medial axis model. *Journal of Vision*, 19(6), 6. <https://doi.org/10.1167/19.6.6>
- Bankieris, K. R., Bejjanki, V. R., & Aslin, R. N. (2017). Sensory cue-combination in the context of newly learned categories. *Scientific Reports*, 7(1), 10890. <https://doi.org/10.1038/s41598-017-11341-7>
- Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and Brain Sciences*, 22(4), 637-660. <https://doi.org/10.1017/S0140525X99532147>
- Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, 59(1), 617-645. <https://doi.org/10.1146/annurev.psych.59.103006.093639>
- Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, 7(2), 84-91. [https://doi.org/10.1016/S1364-6613\(02\)00029-3](https://doi.org/10.1016/S1364-6613(02)00029-3)
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1-48. <https://doi.org/10.18637/jss.v067.i01>
- Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, 57(3), 217-239. [https://doi.org/10.1016/0010-0277\(95\)00669-X](https://doi.org/10.1016/0010-0277(95)00669-X)
- Beauchamp, M. S. (2005). See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Current Opinion in Neurobiology*, 15(2), 145-153. <https://doi.org/10.1016/j.conb.2005.03.011>
- Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, 41(5), 809-823. [https://doi.org/10.1016/S0896-6273\(04\)00070-4](https://doi.org/10.1016/S0896-6273(04)00070-4)
- Beauchamp, M. S., Yasar, N. E., Frye, R. E., & Ro, T. (2008). Touch, sound and vision in human superior temporal sulcus. *Neuroimage*, 41(3), 1011-1020. <https://doi.org/10.1016/j.neuroimage.2008.03.015>
- Beauchamp, M. S., Yasar, N. E., Kishan, N., & Ro, T. (2007). Human MST but not MT responds to tactile stimulation. *Journal of Neuroscience*, 27(31), 8261-8267. <https://doi.org/10.1523/JNEUROSCI.0754-07.2007>
- Beierholm, U., Rohe, T., Ferrari, A., Stegle, O., & Noppeney, U. (2020). Using the past to estimate sensory uncertainty. *Elife*, 9, e54172. <https://doi.org/10.7554/eLife.54172>

- Berlin, B., & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. University of California Press. <https://doi.org/10.1017/S0022226700002966>
- Bernard, C., Kronland-Martinet, R., Fery, M., Ystad, S., & Thoret, E. (2022). Tactile perception of auditory roughness. *JASA Express Letters*, 2(12). <https://doi.org/10.1121/10.0016603>
- Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2), 115–147. <https://doi.org/10.1037/0033-295X.94.2.115>
- Binder, J. R. (2016). In defense of abstract conceptual representations. *Psychonomic Bulletin & Review*, 23(4), 1096–1108. <https://doi.org/10.3758/s13423-015-0909-1>
- Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, 14(10), 693–707. <https://doi.org/10.1038/nrn3565>
- Bizley, J. K., Maddox, R. K., & Lee, A. K. (2016). Defining auditory-visual objects: behavioral tests and physiological mechanisms. *Trends in Neurosciences*, 39(2), 74–85. <https://doi.org/10.1016/j.tins.2015.12.007>
- Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annu. Rev. Psychol.*, 58(1), 47–73. <https://doi.org/10.1146/annurev.psych.57.102904.190152>
- Blender Foundation. (2023). *Blender* (Version 3.6) [Computer software]. <https://www.blender.org>
- Bones, O., Cox, T. J., & Davies, W. J. (2018). Sound categories: Category formation and evidence-based taxonomies. *Frontiers in Psychology*, 9, 1277. <https://doi.org/10.3389/fpsyg.2018.01277>
- Borg, I., Groenen, P. J., Mair, P., Borg, I., Groenen, P. J., & Mair, P. (2012). Variants of different MDS models. *Applied Multidimensional Scaling*, (pp. 37–47). Springer Briefs in Statistics. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-31848-1_5
- Botvinick, M., & Cohen, J. (1998). Rubber hands ‘feel’ touch that eyes see. *Nature*, 391(6669), 756–756. <https://doi.org/10.1038/35784>
- Bracci, S., Ritchie, J. B., & de Beeck, H. O. (2017). On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia*, 105, 153–164. <https://doi.org/10.1016/j.neuropsychologia.2017.06.010>
- Braunstein, M. L. (1990). Structure from motion. In *Human performance models for computer-aided engineering* (pp. 89–105). Academic Press. <https://doi.org/10.1016/B978-0-12-236530-0.50012-1>
- Brefczynski-Lewis, J. A., & Lewis, J. W. (2017). Auditory object perception: A neurobiological model and prospective review. *Neuropsychologia*, 105, 223–242. <https://doi.org/10.1016/j.neuropsychologia.2017.04.034>

- Bregman, A. S. (1984). Auditory scene analysis. In *Proceedings of the 7th International Conference on Pattern Recognition* (pp. 168-175).
- Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. MIT press. <https://doi.org/10.7551/mitpress/1486.001.0001>
- Bregman, A. S. (2017). Asking the “what for” question in auditory perception. In *Perceptual Organization* (pp. 99-118). Routledge.
- Bregman, A. S. (1993). Auditory scene analysis: Hearing in complex environments. In S. McAdams & E. Bigand (Eds.), *Thinking in sound: The cognitive psychology of human audition* (pp. 10–36). Clarendon Press/Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198522577.003.0002>
- Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, 8, e9414. <https://doi.org/10.7717/peerj.9414>
- Briscoe, R. (2020). Colour categorization and categorical perception. *The Routledge handbook of philosophy of colour*, 456-474. <https://doi.org/10.4324/9781351048521-35>
- Broadbent, H., Osborne, T., Kirkham, N., & Mareschal, D. (2020). Touch and look: The role of visual-haptic cues for categorical learning in primary school children. *Infant and Child Development*, 29(2), e2168. <https://doi.org/10.1002/icd.2168>
- Broadbent, H., Osborne, T., Mareschal, D., & Kirkham, N. (2020). Are two cues always better than one? The role of multiple intra-sensory cues compared to multi-cross-sensory cues in children's incidental category learning. *Cognition*, 199, 104202. <https://doi.org/10.1016/j.cognition.2020.104202>
- Brunel, L., Goldstone, R. L., Vallet, G., Riou, B., & Versace, R. (2013). When seeing a dog activates the bark. *Experimental Psychology*, 60(2), 100-112. <https://doi.org/10.1027/1618-3169/a000176>
- Bruner, E. M. (1956). Primary Group Experience and the Processes of Acculturation 1. *American Anthropologist*, 58(4), 605-623. <https://doi.org/10.1525/aa.1956.58.4.02a00030>
- Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, 145(11), 1029–1052. <https://doi.org/10.1037/bul0000209>
- Bullot, N. J., & Égré, P. (2010). Objects and sound perception. *Review of Philosophy and Psychology*, 1(1), 5-17. <https://doi.org/10.1007/s13164-009-0006-3>
- Bülthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain?. *Cerebral Cortex*, 5(3), 247-260. <https://doi.org/10.1093/cercor/5.3.247>

- Bülthoff, I., & Newell, F. (2004). Categorical perception of sex occurs in familiar but not unfamiliar faces. *Visual Cognition*, 11(7), 823-855. <https://doi.org/10.1080/13506280444000012>
- Bülthoff, I., & Newell, F. N. (2006). The role of familiarity in the recognition of static and dynamic objects. *Progress in Brain Research*, 154, 315-325. [https://doi.org/10.1016/S0079-6123\(06\)54017-8](https://doi.org/10.1016/S0079-6123(06)54017-8)
- Bülthoff, I., & Newell, F. N. (2015). Distinctive voices enhance the visual recognition of unfamiliar faces. *Cognition*, 137, 9-21. <https://doi.org/10.1016/j.cognition.2014.12.006>
- Burge, J., Girshick, A. R., & Banks, M. S. (2010). Visual-haptic adaptation is determined by relative reliability. *Journal of Neuroscience*, 30(22), 7714-7721. <https://doi.org/10.1523/JNEUROSCI.6427-09.2010>
- Burton, H., Snyder, A. Z., Diamond, J. B., & Raichle, M. E. (2002). Adaptive changes in early and late blind: a fMRI study of verb generation to heard nouns. *Journal of Neurophysiology*, 88(6), 3359-3371. <https://doi.org/10.1152/jn.00129.2002>
- Calder, A. J., Young, A. W., Perrett, D. I., Etcoff, N. L., & Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition*, 3(2), 81-118. <https://doi.org/10.1080/713756735>
- Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1), 1-27. <https://doi.org/10.1080/03610927408827101>
- Calvert, G. A., & Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *Journal of Physiology-Paris*, 98(1-3), 191-205. <https://doi.org/10.1016/j.jphysparis.2004.03.018>
- Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., ... & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, 276(5312), 593-596. <https://doi.org/10.1126/science.276.5312.593>
- Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage*, 14(2), 427-438. <https://doi.org/10.1006/nimg.2001.0812>
- Camponogara, I., & Volcic, R. (2022). Visual uncertainty unveils the distinct role of haptic cues in multisensory grasping. *eneuro*, 9(3). <https://doi.org/10.1523/ENEURO.0079-22.2022>
- Camus, T., Hommel, B., Brunel, L., & Brouillet, T. (2018). From anticipation to integration: the role of integrated action-effects in building sensorimotor contingencies. *Psychonomic Bulletin & Review*, 25(3), 1059-1065. <https://doi.org/10.3758/s13423-017-1308-6>

- Canits, I., Pecher, D., & Zeelenberg, R. (2018). Effects of grasp compatibility on long-term memory for objects. *Acta Psychologica* (182), 65–74. <https://doi.org/10.1016/j.actpsy.2017.11.009>
- Cantwell, G., Crossley, M. J., & Ashby, F. G. (2015). Multiple stages of learning in perceptual categorization: Evidence and neurocomputational theory. *Psychonomic Bulletin & Review*, 22(6), 1598-1613. <https://doi.org/10.3758/s13423-015-0827-2>
- Cao, Y., Summerfield, C., Park, H., Giordano, B. L., & Kayser, C. (2019). Causal inference in the multisensory brain. *Neuron*, 102(5), 1076-1087. <https://doi.org/10.1016/j.neuron.2019.03.043>
- Cappe, C., Thut, G., Romei, V., & Murray, M. M. (2010). Auditory–visual multisensory interactions in humans: timing, topography, directionality, and sources. *Journal of Neuroscience*, 30(38), 12572-12580. <https://doi.org/10.1523/JNEUROSCI.1099-10.2010>
- Carvalho, P. F., & Goldstone, R. L. (2014a). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, 42(3), 481-495. (2014). <https://doi.org/10.3758/s13421-013-0371-0>
- Carvalho, P. F., & Goldstone, R. L. (2014b). Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology*, 5, 936 - 947. <https://doi.org/10.3389/fpsyg.2014.00936>
- Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, 22(1), 281-288. <https://doi.org/10.3758/s13423-014-0676-4>
- Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(11), 1699. <https://doi.org/10.1037/xlm0000406>
- Carvalho, P. F., & Goldstone, R. L. (2019). When does interleaving practice improve learning? In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge Handbook of Cognition and Education* (pp. 411–436). Cambridge University Press. <https://doi.org/10.1017/9781108235631.017>
- Carvalho, P. F., Chen, C. H., & Yu, C. (2021). The distributional properties of exemplars affect category learning and generalization. *Scientific Reports*, 11(1), 11263. <https://doi.org/10.1038/s41598-021-90743-0>
- Casale, M. B., & Ashby, F. G. (2008). A role for the perceptual representation memory system in category learning. *Perception & Psychophysics*, 70(6), 983-999. <https://doi.org/10.3758/PP.70.6.983>
- Casey, S. J., and Newell, F. N. (2007). Are representations of unfamiliar faces independent of encoding modality? *Neuropsychologia* 45, 506–513. <https://doi.org/10.1016/j.neuropsychologia.2006.02.011>

- Chan, J. S., Simões-Franklin, C., Garavan, H., & Newell, F. N. (2010). Static images of novel, moveable objects learned through touch activate visual area hMT+. *NeuroImage*, 49(2), 1708-1716. <https://doi.org/10.1016/j.neuroimage.2009.09.068>
- Chandrasekaran, B., Yi, H. G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, 21(2), 488-495. <https://doi.org/10.3758/s13423-013-0501-5>
- Chen, Y. C., & 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. <https://doi.org/10.1016/j.cognition.2009.10.012>
- Chen, Y. C., & 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. <https://doi.org/10.1037/a0024329>
- Chen, Y. C., & Spence, C. (2017). Assessing the role of the 'unity assumption' on multisensory integration: A review. *Frontiers in Psychology*, 8, 445. <https://doi.org/10.3389/fpsyg.2017.00445>
- Chiou, R., & Ralph, M. A. L. (2019). Unveiling the dynamic interplay between the hub-and-spoke-components of the brain's semantic system and its impact on human behaviour. *NeuroImage*, 199, 114-126. <https://doi.org/10.1016/j.neuroimage.2019.05.059>
- Choi, I., Lee, J. Y., & Lee, S. H. (2018). Bottom-up and top-down modulation of multisensory integration. *Current Opinion in Neurobiology*, 52, 115-122. <https://doi.org/10.1016/j.conb.2018.05.002>
- Chuang, L. L., Vuong, Q. C., & Bülthoff, H. H. (2012). Learned non-rigid object motion is a view-invariant cue to recognizing novel objects. *Frontiers in Computational Neuroscience*, 6, 26-34. <https://doi.org/10.3389/fncom.2012.00026>
- Chung, Y. H., Schurgin, M. W., & Brady, T. F. (2023). The role of motion in visual working memory for dynamic stimuli: More lagged but more precise representations of moving objects. *Attention, Perception, & Psychophysics*, 85(5), 1387-1397. <https://doi.org/10.3758/s13414-022-02635-8>
- Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J., & Charest, I. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *Neuroimage*, 194, 12-24. <https://doi.org/10.1016/j.neuroimage.2019.03.031>
- Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, 16(2), 409-412. <https://doi.org/10.3758/BF03203962>

- Colonus, H., & Diederich, A. (2020). Formal models and quantitative measures of multisensory integration: a selective overview. *European Journal of Neuroscience*, 51(5), 1161-1178. <https://doi.org/10.1111/ejn.13813>
- Colunga, E., & Smith, L. B. (2008). Knowledge embedded in process: the self-organization of skilled noun learning. *Developmental science*, 11(2), 195–203. <https://doi.org/10.1111/j.1467-7687.2007.00665.x>
- Cooke, T., Jäkel, F., Wallraven, C., & Bülthoff, H. H. (2007). Multimodal similarity and categorization of novel, three-dimensional objects. *Neuropsychologia*, 45(3), 484-495. <https://doi.org/10.1016/j.neuropsychologia.2006.02.009>
- Cooke, T., Kannengiesser, S., Wallraven, C., & Bülthoff, H. H. (2006). Object feature validation using visual and haptic similarity ratings. *ACM Transactions on Applied Perception (TAP)*, 3(3), 239-261. <https://doi.org/10.1145/1166087.1166093>
- Coste, A., Bardy, B. G., Janaqi, S., Słowiński, P., Tsaneva-Atanasova, K., Goupil, J. L., & Marin, L. (2021). Decoding identity from motion: how motor similarities colour our perception of self and others. *Psychological Research*, 85(2), 509–519. <https://doi.org/10.1007/s00426-020-01290-8>
- Cousineau, D. (2004). Merging race models and adaptive networks: A parallel race network. *Psychonomic Bulletin & Review*, 11(5), 807-825. <https://doi.org/10.3758/BF03196707>
- Cox, T. F. (2001). Multidimensional scaling used in multivariate statistical process control. *Journal of Applied Statistics*, 28(3-4), 365-378. <https://doi.org/10.1080/02664760120034108>
- Craddock, M., & Lawson, R. (2009). The effects of size changes on haptic object recognition. *Attention, Perception, & Psychophysics*, 71(4), 910–923. <https://doi.org/10.3758/APP.71.4.910>
- Cutting, J. E., & Kozlowski, L. T. (1977). Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, 9, 353-356. <https://doi.org/10.3758/BF03337021>
- Cutzu F., Edelman S. (1994) Canonical views in object representation and recognition. *Vision Research* (34), 3037-3056. [https://doi.org/10.1016/0042-6989\(94\)90277-1](https://doi.org/10.1016/0042-6989(94)90277-1)
- Cutzu, F., & Edelman, S. (1998). Representation of object similarity in human vision: psychophysics and a computational model. *Vision Research*, 38(15-16), 2229-2257. [https://doi.org/10.1016/S0042-6989\(97\)00186-7](https://doi.org/10.1016/S0042-6989(97)00186-7)
- Davies, W. J., Adams, M. D., Bruce, N. S., Cain, R., Carlyle, A., Cusack, P., ... & Poxon, J. (2013). Perception of soundscapes: An interdisciplinary approach. *Applied Acoustics*, 74(2), 224-231. <https://doi.org/10.1016/j.apacoust.2012.05.010>

- de Beeck, H. O., Wagemans, J., & Vogels, R. (2003). The Effect of Category Learning on the Representation of Shape: Dimensions Can Be Biased but Not Differentiated. *Journal of Experimental Psychology: General*, 132(4), 491–511. <https://doi.org/10.1037/0096-3445.132.4.491>.
- de Beeck, H. P. O., Torfs, K., & Wagemans, J. (2008). Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *Journal of Neuroscience*, 28(40), 10111–10123. <https://doi.org/10.1523/JNEUROSCI.2511-08.2008>
- de Beeck, H. P. O., Torfs, K., & Wagemans, J. (2008). Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *Journal of Neuroscience*, 28(40), 10111–10123. <https://doi.org/10.1523/JNEUROSCI.2511-08.2008>
- De Meo, R., Murray, M. M., Clarke, S., & Matusz, P. J. (2015). Top-down control and early multisensory processes: chicken vs. egg. *Frontiers in integrative neuroscience*, 9, 17–13. <https://doi.org/10.3389/fnint.2015.00017>
- Deng, W. S., & Sloutsky, V. M. (2015). The development of categorization: effects of classification and inference training on category representation. *Developmental Psychology*, 51(3), 392–405. <https://doi.org/10.1037/a0038749>
- Deshpande, G., Hu, X., Lacey, S., Stilla, R., & Sathian, K. (2010). Object familiarity modulates effective connectivity during haptic shape perception. *Neuroimage*, 49(3), 1991–2000. <https://doi.org/10.1016/j.neuroimage.2009.08.052>
- Diaconescu, A. O., Hasher, L., & McIntosh, A. R. (2013). Visual dominance and multisensory integration changes with age. *Neuroimage*, 65, 152–166. <https://doi.org/10.1016/j.neuroimage.2012.09.057>
- DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition?. *Neuron*, 73(3), 415–434. <https://doi.org/10.1016/j.neuron.2012.01.010>
- Doehrmann, O., & Naumer, M. J. (2008). Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. *Brain Research*, 1242, 136–150. <https://doi.org/10.1016/j.brainres.2008.03.071>
- Doehrmann, O., Naumer, M. J., Volz, S., Kaiser, J., & Altmann, C. F. (2008). Probing category selectivity for environmental sounds in the human auditory brain. *Neuropsychologia*, 46(11), 2776–2786. <https://doi.org/10.1016/j.neuropsychologia.2008.05.011>
- Donohue, S. E., & Woldorff, M. G. (2011). The Cross-Modal Spread of Attention Reveals Differential Temporal Dynamics for Synchronous versus Asynchronous Multisensory Events. *Journal of Neuroscience*, 31(22), 7982–7991. <https://doi.org/10.1523/JNEUROSCI.5298-10.2011>

- Doshier, B. A., & Lu, Z. L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences*, 95(23), 13988-13993. <https://doi.org/10.1073/pnas.95.23.13988>
- Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, 293(5539), 2470-2473. <https://doi.org/10.1126/science.1063414>
- Duarte, S. E., Ghetti, S., & Geng, J. J. (2023). Object memory is multisensory: Task-irrelevant sounds improve recollection. *Psychonomic Bulletin & Review*, 30(2), 652–665. <https://doi.org/10.3758/s13423-022-02182-1>
- Duarte, S. E., Yonelinas, A. P., Ghetti, S., & Geng, J. J. (2025). Multisensory processing impacts memory for objects and their sources. *Memory & Cognition*, 53(2), 646-665. <https://doi.org/10.3758/s13421-024-01592-x>
- Dubois, D. (2000). Categories as acts of meaning: The case of categories in olfaction and audition. *Cognitive Science Quarterly*, 1(1), 35-68.
- Dubois, D., Guastavino, C., & Raimbault, M. (2006). A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories. *Acta Acustica United with Acustica*, 92(6), 865-874.
- Easton, R. D., Greene, A. J., & Srinivas, K. (1997). Transfer between vision and haptics: Memory for 2-D patterns and 3-D objects. *Psychonomic Bulletin & Review*, 4(3), 403-410. <https://doi.org/10.3758/BF03210801>
- Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, 21(4), 449-467. <https://doi.org/10.1017/S0140525X98001253>
- Edelman, S. (1999). *Representation and recognition in vision*. MIT press.
- Edelman, S., & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, 32(12), 2385-2400. [https://doi.org/10.1016/0042-6989\(92\)90102-O](https://doi.org/10.1016/0042-6989(92)90102-O)
- Edmunds, C. E. R., Inkster, A. B., Jones, P. M., Milton, F., & Wills, A. J. (2020). Absence of cross-modality analogical transfer in perceptual categorization. *Open Journal of Experimental Psychology and Neuroscience*, 1, 3–13. <https://doi.org/10.46221/ojepn.2020.8639>.
- Enge, A., Süß, F., & Rahman, R. A. (2023). Instant effects of semantic information on visual perception. *Journal of Neuroscience*, 43(26), 4896-4906. <https://doi.org/10.1523/JNEUROSCI.2038-22.2023>

- Engel, L. R., Frum, C., Puce, A., Walker, N. A., & Lewis, J. W. (2009). Different categories of living and non-living sound-sources activate distinct cortical networks. *Neuroimage*, 47(4), 1778-1791. <https://doi.org/10.1016/j.neuroimage.2009.05.041>
- Erdogan, G., Yildirim, I., & Jacobs, R. A. (2014). Transfer of object shape knowledge across visual and haptic modalities. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36(36). <http://dx.doi.org/10.13140/2.1.1513.9206>
- Ernst, M. O. (2006). A Bayesian view on multimodal cue integration. *Human body perception from the inside out*, 131, 105-131.
- Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, 7(5), 7-7. <https://doi.org/10.1167/7.5.7>
- Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870), 429–433. <https://doi.org/10.1038/415429a>
- Ernst, M. O., & Bühlhoff, H. H. (2004). Merging the senses into a robust percept. *Trends in cognitive sciences*, 8(4), 162-169. <https://doi.org/10.1016/j.tics.2004.02.002>
- Ernst, M. O., Lange, C., & Newell, F. N. (2007). Multisensory recognition of actively explored objects. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 61(3), 242–253. <https://doi.org/10.1037/cjep2007025>
- Fairhurst, M. T., Travers, E., Hayward, V., & Deroy, O. (2018). Confidence is higher in touch than in vision in cases of perceptual ambiguity. *Scientific reports*, 8(1), 15604. <https://doi.org/10.1038/s41598-018-34052-z>
- Fang, L., Müller, T., Pescara, E., Fischer, N., Huang, Y., & Beigl, M. (2023). Investigating passive haptic learning of piano songs using three tactile sensations of vibration, stroking and tapping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(3), 1-19. <https://doi.org/10.1145/3610899>
- Fechner, G. T. (1948). Elements of psychophysics. In *Readings in the History of Psychology* (pp. 206–213). Appleton-Century-Crofts. <https://doi.org/10.1037/11304-026>
- Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, 407(6804), 630-633. <https://doi.org/10.1038/35036586>
- Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1), 1-47. <https://doi.org/10.1093/cercor/1.1.1-a>

- Feng, G., Gan, Z., Yi, H. G., Ell, S. W., Roark, C. L., Wang, S., ... & Chandrasekaran, B. (2021). Neural dynamics underlying the acquisition of distinct auditory category structures. *Neuroimage*, 244, 118565. <https://doi.org/10.1016/j.neuroimage.2021.118565>
- Fetsch, C. R., Turner, A. H., DeAngelis, G. C., & Angelaki, D. E. (2009). Dynamic reweighting of visual and vestibular cues during self-motion perception. *Journal of Neuroscience*, 29(49), 15601-15612. <https://doi.org/10.1523/JNEUROSCI.2574-09.2009>
- Fleming, R. W. (2017). Material perception. *Annual Review of Vision Science*, 3(1), 365-388. <https://doi.org/10.1146/annurev-vision-102016-061429>
- Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of Vision*, 13(8), 9-9. <https://doi.org/10.1167/13.8.9>
- Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of Vision*, 13(8), 9-9. <https://doi.org/10.1167/13.8.9>
- Fodor JA. (1983). *The modularity of mind: an essay on faculty psychology*. Cambridge, MA: MIT Press. <https://doi.org/10.2307/2184717>
- Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2012). How category learning affects object representations: not all morphospaces stretch alike. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(4), 807–820. <https://doi.org/10.1037/a0025836>.
- Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex*, 23(4), 814–823. <https://doi.org/10.1093/cercor/bhs067>
- Folstein, J. R., Palmeri, T. J., Van Gulick, A. E., & Gauthier, I. (2015). Category learning stretches neural representations in visual cortex. *Current Directions in Psychological Science*, 24(1), 17-23. <https://doi.org/10.1177/0963721414550707>
- Forster, B., Cavina-Pratesi, C., Aglioti, S. M., & Berlucchi, G. (2002). Redundant target effect and intersensory facilitation from visual-tactile interactions in simple reaction time. *Experimental Brain Research*, 143(4), 480-487. <https://doi.org/10.1007/s00221-002-1017-9>
- Foxe, J. J., & Schroeder, C. E. (2005). The case for feedforward multisensory convergence during early cortical processing. *Neuroreport*, 16(5), 419-423. <https://doi.org/10.1097/00001756-200504040-00001>
- Foxe, J. J., Wylie, G. R., Martinez, A., Schroeder, C. E., Javitt, D. C., Guilfoyle, D., ... & Murray, M. M. (2002). Auditory-somatosensory multisensory processing in auditory association cortex: an fMRI study. *Journal of Neurophysiology*, 88(1), 540-543. <https://doi.org/10.1152/jn.2002.88.1.540>

- Franzen, L., Delis, I., De Sousa, G., Kayser, C., & Philiastides, M. G. (2020). Auditory information enhances post-sensory visual evidence during rapid multisensory decision-making. *Nature Communications*, 11(1), 5440. <https://doi.org/10.1038/s41467-020-19306-7>
- Freedman, D. J., & Miller, E. K. (2008). Neural mechanisms of visual categorization: insights from neurophysiology. *Neuroscience & Biobehavioral Reviews*, 32(2), 311-329. <https://doi.org/10.1016/j.neubiorev.2007.07.011>
- Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502), 312-316. <https://doi.org/10.1126/science.291.5502.312>
- Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2002). Visual categorization and the primate prefrontal cortex: neurophysiology and behavior. *Journal of Neurophysiology*, 88(2), 929-941. <https://doi.org/10.1152/jn.2002.88.2.929>
- Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, 23(12), 5235-5246. <https://doi.org/10.1523/JNEUROSCI.23-12-05235.2003>
- Freiwald, J. P., Göbel, Y., Mostajeran, F., & Steinicke, F. (2020). The cybersickness susceptibility questionnaire: predicting virtual reality tolerance. In *Proceedings of mensch und computer 2020*, 115-118. <https://doi.org/10.1145/3404983.3410022>
- Frissen, I., Kappasov, Z., Huang, K. Y., & Ziat, M. (2023). Humans can sense small numbers of objects in a box by touch alone. *Perception*, 52(11-12), 799-811. <https://doi.org/10.1177/03010066231201960>
- Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456), 815-836. <https://doi.org/10.1098/rstb.2005.1622>
- Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature Reviews Neuroscience*, 11(2), 127-138. <https://doi.org/10.1038/nrn2787>
- Fugate, J. M. (2013). Categorical perception for emotional faces. *Emotion Review*, 5(1), 84-89. <https://doi.org/10.1177/1754073912451350>
- Gabora, L., Rosch, E., & Aerts, D. (2008). Toward an ecological theory of concepts. *Ecological Psychology*, 20(1), 84-116. <https://doi.org/10.1080/10407410701766676>
- Gaissert, N., & Wallraven, C. (2012). Categorizing natural objects: a comparison of the visual and the haptic modalities. *Experimental Brain Research*, 216(1), 123-134. <https://doi.org/10.1007/s00221-011-2916-4>

- Gaißert, N., Bülthoff, H. H., & Wallraven, C. (2011). Similarity and categorization: From vision to touch. *Acta Psychologica*, 138(1), 219-230. <https://doi.org/10.1016/j.actpsy.2011.06.007>
- Gaissert, N., Wallraven, C., & Bülthoff, H. H. (2010). Visual and haptic perceptual spaces show high similarity in humans. *Journal of Vision*, 10(11), 2. <https://doi.org/10.1167/10.11.2>
- Gaißert, N., Waterkamp, S., Fleming, R. W., & Bülthoff, I. (2012). Haptic categorical perception of shape. *PloS one*, 7(8), e43062. <https://doi.org/10.1371/journal.pone.0043062>
- Gallace, A., & Spence, C. (2014). *In Touch with the Future: The Sense of Touch from Cognitive Neuroscience to Virtual Reality*. Oxford University Press.
- Ganel, T., & Goshen-Gottstein, Y. (2004). Effects of familiarity on the perceptual integrality of the identity and expression of faces: the parallel-route hypothesis revisited. *Journal of Experimental Psychology: Human Perception and Performance*, 30(3), 583–597. <https://doi.org/10.1037/0096-1523.30.3.583>
- Garner, W. R. (2014). The processing of information and structure. *Psychology Press*. <https://doi.org/10.4324/9781315802862>
- Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology*, 1(3), 225-241. [https://doi.org/10.1016/0010-0285\(70\)90016-2](https://doi.org/10.1016/0010-0285(70)90016-2)
- Gau, R., Bazin, P. L., Trampel, R., Turner, R., & Noppeney, U. (2020). Resolving multisensory and attentional influences across cortical depth in sensory cortices. *elife*, 9, e46856. <https://doi.org/10.7554/eLife.46856>
- Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3(2), 191-197. <https://doi.org/10.1038/72140>
- Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience*, 2(6), 568-573. <https://doi.org/10.1038/9224>
- Ge, Y., Li, F., Li, X., & Li, W. (2021). What is the mechanism underlying the interleaving effect in category induction: An eye-tracking and behavioral study. *Frontiers in Psychology*, 12, 770885. <https://doi.org/10.3389/fpsyg.2021.770885>
- Gentaz, E., Baud-Bovy, G., & Luyat, M. (2008). The haptic perception of spatial orientations. *Experimental Brain Research*, 187(3), 331-348. <https://doi.org/10.1007/s00221-008-1382-0>
- Gepshtein, S., & Banks, M. S. (2003). Viewing geometry determines how vision and haptics combine in size perception. *Current Biology*, 13(6), 483–488. [https://doi.org/10.1016/S0960-9822\(03\)00133-7](https://doi.org/10.1016/S0960-9822(03)00133-7)

- Gepshtein, S., Burge, J., Ernst, M. O., & Banks, M. S. (2005). The combination of vision and touch: Measurement of weights for adults and children. In *ACM International Conference Proceeding Series; Vol. 119. Proceedings of the 2005 symposium on Applied perception in graphics and visualization* (pp. 157-157). <https://doi.org/10.1167/5.11.7>
- Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory?. *Trends in cognitive sciences*, 10(6), 278-285. <https://doi.org/10.1016/j.tics.2006.04.008>
- 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. <https://doi.org/10.1162/089892999563544>
- Gibson, J. J. (1962). Observations on active touch. *Psychological Review*, 69(6), 477–491. <https://doi.org/10.1037/h0046962>
- Girondini, M., Bertoni, T., Montanaro, M., Serino, A., & Gallace, A. (2025). Spatial sensorimotor mismatch increases the excitability of the primary somatosensory cortex: Insight from an EEG-Virtual Reality Study. *Cortex*. <https://doi.org/10.1016/j.cortex.2025.07.010>
- Girondini, M., Montanaro, M., & Gallace, A. (2024). Exploring virtual reality object perception following sensory-motor interactions with different visuo-haptic collider properties. *Scientific reports*, 14(1), 10011. <https://doi.org/10.1038/s41598-024-59570-x>
- Gleiss, S., & Kayser, C. (2014). Oscillatory mechanisms underlying the enhancement of visual motion perception by multisensory congruency. *Neuropsychologia*, 53, 84-93. <https://doi.org/10.1016/j.neuropsychologia.2013.11.005>
- Goldstone, R. L. (1994a). The role of similarity in categorization: Providing a groundwork. *Cognition*, 52(2), 125-157. [https://doi.org/10.1016/0010-0277\(94\)90065-5](https://doi.org/10.1016/0010-0277(94)90065-5)
- Goldstone, R. L. (1994b). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123(2), 178–200. <https://doi.org/10.1037/0096-3445.123.2.178>
- Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, 49(1), 585-612. <https://doi.org/10.1146/annurev.psych.49.1.585>
- Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, 65(2-3), 231-262. [https://doi.org/10.1016/S0010-0277\(97\)00047-4](https://doi.org/10.1016/S0010-0277(97)00047-4)
- Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley interdisciplinary reviews. Cognitive Science*, 1(1), 69–78. <https://doi.org/10.1002/wcs.26>

- Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, 130(1), 116–139. <https://doi.org/10.1037/0096-3445.130.1.116>.
- Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, 78(1), 27–43. [https://doi.org/10.1016/s0010-0277\(00\)00099-8](https://doi.org/10.1016/s0010-0277(00)00099-8)
- Goldstone, R. L., Steyvers, M., & Larimer, K. (2019). Categorical perception of novel dimensions. In *Proceedings of the eighteenth annual conference of the cognitive science society* (pp. 243-248). Routledge.
- Gonçalves, G., Coelho, H., Monteiro, P., Melo, M., & Bessa, M. (2022). Systematic review of comparative studies of the impact of realism in immersive virtual experiences. *ACM Computing Surveys*, 55(6), 1-36. <https://doi.org/10.1145/3533377>
- Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1), 20-25. [https://doi.org/10.1016/0166-2236\(92\)90344-8](https://doi.org/10.1016/0166-2236(92)90344-8)
- Gori, M., Price, S., Newell, F. N., Berthouze, N., & Volpe, G. (2022). Multisensory Perception and Learning: Linking Pedagogy, Psychophysics, and Human–Computer Interaction. *Multisensory Research*, 35(4), 335-366. <https://doi.org/10.1163/22134808-bja10072>
- Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object?. *Nature Reviews Neuroscience*, 5(11), 887-892. <https://doi.org/10.1038/nrn1538>
- Grill-Spector, K. (2003). The neural basis of object perception. *Current Opinion in Neurobiology*, 13(2), 159-166. [https://doi.org/10.1016/S0959-4388\(03\)00040-0](https://doi.org/10.1016/S0959-4388(03)00040-0).
- Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, 41(10-11), 1409-1422. [https://doi.org/10.1016/S0042-6989\(01\)00073-6](https://doi.org/10.1016/S0042-6989(01)00073-6)
- Grill-Spector, K., Kushnir, T., Hendler, T., & Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nature Neuroscience*, 3(8), 837-843. <https://doi.org/10.1038/77754>
- Grootswagers, T. (2020). A primer on running human behavioural experiments online. *Behavior Research Methods*, 52(6), 2283-2286. <https://doi.org/10.3758/s13428-020-01395-3>
- Grossman, E. D., & Blake, R. (1999). Perception of coherent motion, biological motion and form-from-motion under dim-light conditions. *Vision Research*, 39(22), 3721-3727. [https://doi.org/10.1016/S0042-6989\(99\)00084-X](https://doi.org/10.1016/S0042-6989(99)00084-X)

- Grossman, M., Smith, E. E., Koenig, P., Glosser, G., DeVita, C., Moore, P., & McMillan, C. (2002). The neural basis for categorization in semantic memory. *Neuroimage*, 17(3), 1549-1561. <https://doi.org/10.1006/nimg.2002.1273>
- Grzywacz, N. M., & Hildreth, E. C. (1987). Incremental rigidity scheme for recovering structure from motion: position-based versus velocity-based formulations. *Journal of the Optical Society of America A*, 4(3), 503-518. <https://doi.org/10.1364/josaa.4.000503>
- Guastavino, C. (2007). Categorization of environmental sounds. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 61(1), 54-63. <http://dx.doi.org/10.1037/cjep2007006>
- Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *The Journal of the Acoustical Society of America*, 100(2), 1111-1121. <https://doi.org/10.1121/1.416296>
- Gureckis, T. M., & Goldstone, R. L. (2008). The effect of the internal structure of categories on perception. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (10), 843– 848.
- Haag, S. (2011). Effects of vision and haptics on categorizing common objects. *Cognitive Processing*, 12, 33-39. <https://doi.org/10.1007/s10339-010-0369-5>
- Hahn, U. E., & Ramscar, M. E. (2001). *Similarity and categorization*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198506287.001.0001>
- Hampton, J. A. (2006). Concepts as prototypes. *Psychology of Learning and Motivation*, 46, 79-113. [https://doi.org/10.1016/S0079-7421\(06\)46003-5](https://doi.org/10.1016/S0079-7421(06)46003-5)
- Hanada, G. M., Ahveninen, J., Calabro, F. J., Yengo-Kahn, A., & Vaina, L. M. (2019). Cross-modal cue effects in motion processing. *Multisensory Research*, 32(1), 45-65. <https://doi.org/10.1163/22134808-20181313>
- Hanley, J. R., & Roberson, D. (2011). Categorical perception effects reflect differences in typicality on within-category trials. *Psychonomic Bulletin & Review*, 18(2), 355–363. <https://doi.org/10.3758/s13423-010-0043-z>
- Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview. <http://eprints.soton.ac.uk/id/eprint/250386>
- Harnad, S. (2005). Distributed processes, distributed cognizers, and collaborative cognition. *Pragmatics & Cognition*, 13(3), 501-514. <https://doi.org/10.1075/pc.13.3.06har>

- Harrison, C. L., & Gillan, D. J. (2005). The Role of Motion in Object Recognition. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(17), 1625-1629. <https://doi.org/10.1177/154193120504901727>
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425-2430. <https://doi.org/10.1126/science.1063736>
- Hayward, W. G. (1998). Effects of outline shape in object recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 24(2), 427–440. <https://doi.org/10.1037/0096-1523.24.2.427>
- Heald, S. L., Van Hedger, S. C., & Nusbaum, H. C. (2017). Perceptual plasticity for auditory object recognition. *Frontiers in Psychology*, 8, 781- 797. <https://doi.org/10.3389/fpsyg.2017.00781>
- Heald, S. L., Van Hedger, S. C., Veillette, J., Reis, K., Snyder, J. S., & Nusbaum, H. C. (2022). Going beyond rote auditory learning: Neural patterns of generalized auditory learning. *Journal of Cognitive Neuroscience*, 34(3), 425-444. https://doi.org/10.1162/jocn_a_01805
- Heider, F. (1959). On perception, event structure, and the psychological environment. *Psychological Issues*, 1(3), 1–123.
- Heikkilä, J., Alho, K., Hyvönen, H., & Tiippana, K. (2015). Audiovisual semantic congruency during encoding enhances memory performance. *Experimental Psychology*, 62(2), 123–130. <https://doi.org/10.1027/1618-3169/a000279>
- Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6), 1264-1282. <https://doi.org/10.1037/0278-7393.20.6.1264>
- Helbig, H. B., & Ernst, M. O. (2007). Optimal integration of shape information from vision and touch. *Experimental Brain Research*, 179(4), 595–606. <https://doi.org/10.1007/s00221-006-0814-y>
- Helbig, H. B., Ernst, M. O., Ricciardi, E., Pietrini, P., Thielscher, A., Mayer, K. M., ... & Noppeney, U. (2012). The neural mechanisms of reliability weighted integration of shape information from vision and touch. *Neuroimage*, 60(2), 1063-1072. <https://doi.org/10.1016/j.neuroimage.2011.09.072>
- Heller, M. A. (1989). Texture perception in sighted and blind observers. *Perception & Psychophysics*, 45(1), 49-54. <https://doi.org/10.3758/BF03208032>
- Heller, M. A. (1992). Haptic dominance in form perception: vision versus proprioception. *Perception*, 21(5), 655-660. <https://doi.org/10.1068/p210655>

- Henderson, M. M., Serences, J. T., & Rungratsameetaweemana, N. (2025). Dynamic categorization rules alter representations in human visual cortex. *Nature Communications*, 16(1), 3459. <https://doi.org/10.1038/s41467-025-58707-4>
- Henriques, D. Y., & Soechting, J. F. (2003). Bias and sensitivity in the haptic perception of geometry. *Experimental Brain Research*, 150(1), 95-108. <https://doi.org/10.1007/s00221-003-1402-z>
- Hernández-Pérez, R., Cuaya, L. V., Rojas-Hortelano, E., Reyes-Aguilar, A., Concha, L., & De Lafuente, V. (2017). Tactile object categories can be decoded from the parietal and lateral-occipital cortices. *Neuroscience*, 352, 226-235. <https://doi.org/10.1016/j.neuroscience.2017.03.038>
- Hershberger, W. A., & Misceo, G. F. (1996). Touch dominates haptic estimates of discordant visual-haptic size. *Perception & Psychophysics*, 58(7), 1124-1132. <https://doi.org/10.3758/BF03206838>
- Hintzman, D. L. (1986). "Schema Abstraction" in a multiple-trace memory model. *Psychological Review*, 93(4), 411-428. <https://doi.org/10.1037/0033-295X.93.4.411>
- Hirst, R. J., McGovern, D. P., Setti, A., Shams, L., & Newell, F. N. (2020). What you see is what you hear: Twenty years of research using the Sound-Induced Flash Illusion. *Neuroscience & Biobehavioral Reviews*, 118, 759-774. <https://doi.org/10.1016/j.neubiorev.2020.09.006>
- Hirst, R., Roberts, K., Seveso, M. A., O'Dowd, A., Peirce, J., & Newell, F. N. (in review). Delivering Tactile Stimuli via Mobile Browsers: A Method for Remote Multisensory Research. https://doi.org/10.31219/osf.io/y84js_v1
- Hollins, M., & Risner, S. R. (2000). Evidence for the duplex theory of tactile texture perception. *Perception & Psychophysics*, 62(4), 695-705. <https://doi.org/10.3758/BF03206916>
- Hollins, M., Sliman, J., Bensmaïa, S., & Washburn, M. (2001). Vibrotactile adaptation impairs discrimination of fine, but not coarse, textures. *Somatosensory & Motor Research*, 18(4), 253-262. <https://doi.org/10.1080/01421590120089640>
- Holmes, N. P., & Spence, C. (2005). Multisensory integration: space, time and superadditivity. *Current Biology*, 15(18), R762-R764. <https://doi.org/10.1016/j.cub.2005.08.058>
- Horowitz, L. M., & Turan, B. (2008). Prototypes and personal templates: collective wisdom and individual differences. *Psychological Review*, 115(4), 1054-1068. <https://doi.org/10.1037/a0013122>
- Howard, I. P., & Templeton, W. B. (1966). *Human spatial orientation*. John Wiley & Sons.
- Howard, J. H., & Ballas, J. A. (1980). Syntactic and semantic factors in the classification of nonspeech transient patterns. *Perception & Psychophysics*, 28(5), 431-439. <https://doi.org/10.3758/BF03204887>

- Huard, A., Chen, M., & Sra, M. (2022). CardsVR: a two-person VR experience with passive haptic feedback from a deck of playing cards. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* 538-547, <https://doi.org/10.1109/ISMAR55827.2022.00070>
- Hutmacher, F. (2019). Why is there so much more research on vision than on any other sensory modality?. *Frontiers in Psychology*, 10, 481030. <https://doi.org/10.3389/fpsyg.2019.02246>
- Iachini, T., Borghi, A. M., & Senese, V. P. (2008). Categorization and sensorimotor interaction with objects. *Brain and Cognition*, 67(1), 31-43. <https://doi.org/10.1016/j.bandc.2007.11.002>
- Iamshchinina, P., Karapetian, A., Kaiser, D., & Cichy, R. M. (2022). Resolving the time course of visual and auditory object categorization. *Journal of Neurophysiology*, 127(6), 1622-1628. <https://doi.org/10.1152/jn.00515.2021>
- Intraub H. (1980). Presentation rate and the representation of briefly glimpsed pictures in memory. *Journal of Experimental Psychology. Human Learning and Memory*, 6(1), 1–12. <https://doi.org/10.1037/0278-7393.6.1.1>
- Inuggi, A., Domenici, N., Tonelli, A., & Gori, M. (2024). PsySuite: An android application designed to perform multimodal psychophysical testing. *Behavior Research Methods*, 56(8), 8308-8329. <https://doi.org/10.3758/s13428-024-02475-4>
- Jacobs, R. A., & Xu, C. (2019). Can multisensory training aid visual learning? A computational investigation. *Journal of Vision*, 19(11), 1-1. <https://doi.org/10.1167/19.11.1>
- James, T. W., Humphrey, G. K., Gati, J. S., Servos, P., Menon, R. S., & Goodale, M. A. (2002). Haptic study of three-dimensional objects activates extrastriate visual areas. *Neuropsychologia*, 40(10), 1706-1714. [https://doi.org/10.1016/S0028-3932\(02\)00017-9](https://doi.org/10.1016/S0028-3932(02)00017-9)
- James, T. W., Kim, S., & Fisher, J. S. (2007). The neural basis of haptic object processing. *Canadian Journal of Experimental Psychology*, 61(3), 219-229. <https://doi.org/10.1037/cjep2007023>
- Jansen, S. E., Tiest, W. M. B., & Kappers, A. M. (2013). Identifying haptic exploratory procedures by analyzing hand dynamics and contact force. *IEEE Transactions on Haptics*, 6(4), 464-472. <https://doi.org/10.1109/TOH.2013.22>
- Jao, R. J., James, T. W., & James, K. H. (2014). Multisensory convergence of visual and haptic object preference across development. *Neuropsychologia*, 56, 381-392. <https://doi.org/10.1016/j.neuropsychologia.2014.02.009>
- Jastorff, J., Popivanov, I. D., Vogels, R., Vanduffel, W., & Orban, G. A. (2012). Integration of shape and motion cues in biological motion processing in the monkey STS. *Neuroimage*, 60(2), 911-921. <https://doi.org/10.1016/j.neuroimage.2011.12.087>

- Jeannerod, M. (1988). *The neural and behavioural organization of goal-directed movements*. Clarendon Press/Oxford University Press.
- Jiang, J., Wagner, A. D., & Egner, T. (2018). Integrated externally and internally generated task predictions jointly guide cognitive control in prefrontal cortex. *Elife*, 7, e39497. <https://doi.org/10.7554/eLife.39497>
- Jiang, X., Chevillet, M. A., Rauschecker, J. P., & Riesenhuber, M. (2018). Training humans to categorize monkey calls: auditory feature-and category-selective neural tuning changes. *Neuron*, 98(2), 405-416. <https://doi.org/10.1016/j.neuron.2018.03.014>
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14, 201-211. <https://doi.org/10.3758/BF03212378>
- Jones, L. A., & Lederman, S. J. (2006). *Human hand function*. Oxford university press. <https://doi.org/10.1093/acprof:oso/9780195173154.001.0001>
- Jones, M. G., Bokinsky, A., Tretter, T., & Negishi, A. (2005). A comparison of learning with haptic and visual modalities. *Haptics-e, The Electronic Journal of Haptics Research*, 3(6), 1-20. <http://hdl.handle.net/1773/34891>
- Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research*, 47(26), 3286-3297. <https://doi.org/10.1016/j.visres.2007.09.013>
- Kalish, M. L., Lewandowsky, S., & Davies, M. (2005). Error-driven knowledge restructuring in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5), 846-861. <https://doi.org/10.1037/0278-7393.31.5.846>
- Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11), 4302-4311. <https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997>
- Kappers, A. M., & Bergmann Tiest, W. M. (2013). Haptic perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(4), 357-374. <https://doi.org/10.1002/wcs.1238>
- Kappers, A. M., & Bergmann Tiest, W. M. (2014). Influence of shape on the haptic size aftereffect. *PLoS One*, 9(2), e88729. <https://doi.org/10.1371/journal.pone.0088729>
- Kappers, A. M., & Tiest, W. M. B. (2013). Haptic size aftereffects revisited. In *2013 World Haptics Conference (WHC)* (pp. 335-339). IEEE. <https://doi.org/10.1109/WHC.2013.6548431>

- Kassuba, T., Klinge, C., Hölig, C., Röder, B., and Siebner, H. R. (2013). Vision holds a greater share in visuo-haptic object recognition than touch. *Neuroimage* 65, 59–68. <https://doi.org/10.1016/j.neuroimage.2012.09.054>
- Katz, D. (1989). *The world of touch*. Psychology press. <https://doi.org/10.4324/9780203771976>
- Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *Language*, 39(2), 170-210. <https://www.istor.org/stable/411200>
- Kaup, B., Ulrich, R., Bausenhardt, K. M., Bryce, D., Butz, M. V., Dignath, D., ... & Wong, H. Y. (2024). Modal and amodal cognition: an overarching principle in various domains of psychology. *Psychological Research*, 88(2), 307-337. <https://doi.org/10.1007/s00426-023-01878-w>
- Kayser, C., & Shams, L. (2015). Multisensory causal inference in the brain. *PLoS biology*, 13(2), e1002075. <https://doi.org/10.1371/journal.pbio.1002075>
- Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, 3(3), 203-220. https://doi.org/10.1207/s15327108ijap0303_3
- Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol.*, 55(1), 271-304. <https://doi.org/10.1146/annurev.psych.55.090902.142005>
- Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex*, 48(7), 805-825. <https://doi.org/10.1016/j.cortex.2011.04.006>
- Kiltner, K., Groten, R., & Slater, M. (2012). The Sense of Embodiment in Virtual Reality. *Presence: Teleoperators and Virtual Environments*, 21(4), 373–387. https://doi.org/10.1162/PRES_a_00124
- Kim, D., Kim, Y., & Jo, D. (2022). Exploring the effect of virtual environments on passive haptic perception. *Applied Sciences*, 13(1), 299 - 312. <https://doi.org/10.3390/app13010299>
- Kim, S., & James, T. W. (2010). Enhanced effectiveness in visuo-haptic object-selective brain regions with increasing stimulus salience. *Human Brain Mapping*, 31(5), 678–693. <https://doi.org/10.1002/hbm.20897>
- Kitada, R., Hashimoto, T., Kochiyama, T., Kito, T., Okada, T., Matsumura, M., ... & Sadato, N. (2005). Tactile estimation of the roughness of gratings yields a graded response in the human brain: an fMRI study. *Neuroimage*, 25(1), 90-100. <https://doi.org/10.1016/j.neuroimage.2004.11.026>
- Klatzky, R. L., & Lederman, S. J. (1995). Identifying objects from a haptic glance. *Perception & Psychophysics*, 57(8), 1111–1123. <https://doi.org/10.3758/BF03208368>

- Klatzky, R. L., & Lederman, S. J. (2011). Haptic object perception: spatial dimensionality and relation to vision. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1581), 3097-3105. <https://doi.org/10.1098/rstb.2011.0153>
- Klatzky, R. L., & Peck, J. (2011). Please touch: Object properties that invite touch. *IEEE Transactions on Haptics*, 5(2), 139-147. <https://doi.org/10.1109/TOH.2011.54>
- Klatzky, R. L., Lederman, S. J., & Metzger, V. A. (1985). Identifying objects by touch: An "expert system." *Perception & Psychophysics*, 37(4), 299–302. <https://doi.org/10.3758/BF03211351>
- Klatzky, R. L., Lederman, S. J., & Reed, C. (1987). There's more to touch than meets the eye: The salience of object attributes for haptics with and without vision. *Journal of Experimental Psychology: General*, 116(4), 356 - 369. <https://doi.org/10.1037/0096-3445.116.4.356>
- Klatzky, R., & Lederman, S. (2007). Object recognition by touch. In *Blindness and brain plasticity in navigation and object perception* (pp. 197-220). Psychology Press.
- Knight, F. L. C., Longo, M. R., & Bremner, A. J. (2014). Categorical perception of tactile distance. *Cognition*, 131(2), 254-262. <https://doi.org/10.1016/j.cognition.2014.01.005>
- Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12), 712-719. <https://doi.org/10.1016/j.tins.2004.10.007>
- Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, 74(6), 1114–1124. <https://doi.org/10.1016/j.neuron.2012.04.036>
- Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS One*, 2(9), e943. <https://doi.org/10.1371/journal.pone.0000943>
- Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the “enemy of induction”? *Psychological Science*, 19(6), 585-592. <https://doi.org/10.1111/j.1467-9280.2008.02127.x>
- Kost, A. S., Carvalho, P. F., & Goldstone, R. L. (2015). Can you repeat that? The effect of item repetition on interleaved and blocked study. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 37). <https://escholarship.org/uc/item/864370qv>
- Kourtzi Z, Betts LR, Sarkheil P, Welchman AE (2005) Distributed Neural Plasticity for Shape Learning in the Human Visual Cortex. *PLoS Biol* 3(7): e204. <https://doi.org/10.1371/journal.pbio.0030204>
- Kravitz, D. J., & Behrmann, M. (2011). Space-, object-, and feature-based attention interact to organize visual scenes. *Attention, Perception, & Psychophysics*, 73(8), 2434-2447. <https://doi.org/10.3758/s13414-011-0201-z>

- Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, 23(6), 1681–1712. <https://doi.org/10.3758/s13423-016-1049-y>
- Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, 5(1), 3–36. <https://doi.org/10.1080/09540099308915683>
- Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1-27. <https://doi.org/10.1007/BF02289565>
- Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2), 115-129. <https://doi.org/10.1007/BF02289694>
- Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. SAGE Publications, Inc., <https://doi.org/10.4135/9781412985130>
- Kubovy, M., & Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition*, 80(1-2), 97-126. [https://doi.org/10.1016/S0010-0277\(00\)00155-4](https://doi.org/10.1016/S0010-0277(00)00155-4)
- Kuhl, P. K. (1991). Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50(2), 93-107. <https://doi.org/10.3758/BF03212211>
- Kuhl, P. K., & Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception & Psychophysics*, 32(6), 542-550. <https://doi.org/10.3758/BF03204208>
- Kyler, H., & James, K. (2024). The importance of multisensory-motor learning on subsequent visual recognition. *Perception*, 53(9), 597-618. <https://doi.org/10.1177/03010066241258967>
- Labov, W. (1973). The boundaries of words and their meanings. In C.-J. N. Bailey & R. W. Shuy (Eds.), *New Ways of Analyzing Variation in English* (pp. 340-373). Georgetown University Press.
- Lacey, S., & Sathian, K. (2014). Visuo-haptic multisensory object recognition, categorization, and representation. *Frontiers in Psychology*, 5, 730. <https://doi.org/10.3389/fpsyg.2014.00730>
- Lacey, S., & Sathian, K. (2020). Visuo-haptic object perception. *Multisensory Perception*, 157-178. <https://doi.org/10.1016/B978-0-12-812492-5.00007-3>
- Lacey, S., and Campbell, C. (2006). Mental representation in visual/haptic crossmodal memory: evidence from interference effects. *Q. J. Exp. Psychol.* 59, 361–376. <https://doi.org/10.1080/17470210500173232>

- Lacey, S., Flueckiger, P., Stilla, R., Lava, M., & Sathian, K. (2010). Object familiarity modulates the relationship between visual object imagery and haptic shape perception. *Neuroimage*, 49(3), 1977-1990. <https://doi.org/10.1016/j.neuroimage.2009.10.081>
- Lacey, S., Pappas, M., Kreps, A., Lee, K., & Sathian, K. (2009). Perceptual learning of view-independence in visuo-haptic object representations. *Experimental Brain Research*, 198(2), 329-337. <https://doi.org/10.1007/s00221-009-1856-8>
- Lacey, S., Peters, A., and Sathian, K. (2007). Cross-modal object representation is viewpoint-independent. *PLoS One* 2:e890. <https://doi.org/10.1371/journal.pone0000890>
- Lacey, S., Tal, N., Amedi, A., & Sathian, K. (2009). A putative model of multisensory object representation. *Brain Topography*, 21(3), 269-274. <https://doi.org/10.1007/s10548-009-0087-4>
- Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press.
- Lambon Ralph, M. A., Pobric, G., & Jefferies, E. (2009). Conceptual knowledge is underpinned by the temporal pole bilaterally: convergent evidence from rTMS. *Cerebral cortex*, 19(4), 832-838. <https://doi.org/10.1093/cercor/bhn131>
- Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature reviews neuroscience*, 18(1), 42-55. <https://doi.org/10.1038/nrn.2016.150>
- Lambon Ralph, M. A., & Patterson, K. (2008). Generalization and differentiation in semantic memory: insights from semantic dementia. *Annals of the new York Academy of Sciences*, 1124(1), 61-76. <https://doi.org/10.1196/annals.1440.006>
- Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Research*, 35(3), 389-412. [https://doi.org/10.1016/0042-6989\(94\)00176-M](https://doi.org/10.1016/0042-6989(94)00176-M)
- 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. <https://doi.org/10.1007/s00221-004-1913-2>
- Lawson, R. (2009). A comparison of the effects of depth rotation on visual and haptic three-dimensional object recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 35(4), 911–930. <https://doi.org/10.1037/a0015025>

- Lawson, R. (2011). An investigation into the cause of orientation-sensitivity in haptic object recognition. *Seeing and Perceiving*, 24(3), 293-314. <https://doi.org/10.1163/187847511X579052>
- Lawson, R. (2014). Recognizing familiar objects by hand and foot: Haptic shape perception generalizes to inputs from unusual locations and untrained body parts. *Attention, Perception, & Psychophysics*, 76(2), 541-558. <https://doi.org/10.3758/s13414-013-0559-1>
- Lawson, R., & Bracken, S. (2011). Haptic object recognition: How important are depth cues and plane orientation?. *Perception*, 40(5), 576-597. <https://doi.org/10.1068/p6786>
- Lawson, R., Fernandes, A. M., Albuquerque, P. B., & Lacey, S. (2015). Remembering touch: Using interference tasks to study tactile and haptic memory. *Mechanisms of sensory working memory: Attention and performance XXV*, 229–249. <https://doi.org/10.1016/B978-0-12-801371-7.00019-3>
- Le, T. S., Bailly, G., Vezzoli, E., Auvray, M., & Gueorguiev, D. (2024). Tactile sensitivity to missing tones in complex vibrotactile signals. *Journal of neurophysiology*, 132(6), 1943–1954. <https://doi.org/10.1152/jn.00430.2023>
- Lederman, S. J., & Abbott, S. G. (1981). Texture perception: studies of intersensory organization using a discrepancy paradigm, and visual versus tactual psychophysics. *Journal of Experimental Psychology: Human Perception and Performance*, 7(4), 902–915. <https://doi.org/10.1037/0096-1523.7.4.902>
- Lederman, S. J., & Klatzky, R. L. (1987). Hand movements: A window into haptic object recognition. *Cognitive Psychology*, 19(3), 342–368. [https://doi.org/10.1016/0010-0285\(87\)90008-9](https://doi.org/10.1016/0010-0285(87)90008-9)
- Lederman, S. J., & Klatzky, R. L. (1990). Haptic classification of common objects: Knowledge-driven exploration. *Cognitive Psychology*, 22(4), 421-459. [https://doi.org/10.1016/0010-0285\(90\)90009-5](https://doi.org/10.1016/0010-0285(90)90009-5)
- Lederman, S. J., & Klatzky, R. L. (2004). Haptic identification of common objects: Effects of constraining the manual exploration process. *Perception & Psychophysics*, 66(4), 618-628. <https://doi.org/10.3758/BF03194906>
- Lederman, S. J., & Klatzky, R. L. (2009). Haptic perception: A tutorial. *Attention, Perception, & Psychophysics*, 71(7), 1439-1459. <https://doi.org/10.3758/APP.71.7.1439>
- Lee Masson, H., Bulthé, J., Op de Beeck, H. P., & Wallraven, C. (2016). Visual and haptic shape processing in the human brain: unisensory processing, multisensory convergence, and top-down influences. *Cerebral Cortex*, 26(8), 3402-3412. <https://doi.org/10.1093/cercor/bhv170>
- Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Brain research. Cognitive Brain Research*, 24(2), 326–334. <https://doi.org/10.1016/j.cogbrainres.2005.02.005>

- Lenth Russell, V., Paul, B., Maxime, H., Maarten, J., Jonathon, L., Fernando, M., ... & Henrik, S. (2022). emmeans: Estimated marginal means, aka least-squares means. R package version 1.7, 5.
- Leo, F., Sandini, G., & Sciutti, A. (2022). Mental rotation skill shapes haptic exploration strategies. *IEEE Transactions on Haptics*, 15(2), 339-350. <https://doi.org/10.1109/TOH.2022.3162321>
- Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, 4(1), 89–94. <https://doi.org/10.1038/82947>
- Lescroart, M. D., & Biederman, I. (2013). Cortical representation of medial axis structure. *Cerebral Cortex (New York, N.Y. : 1991)*, 23(3), 629–637. <https://doi.org/10.1093/cercor/bhs046>
- Leshinskaya, A., & Caramazza, A. (2016). For a cognitive neuroscience of concepts: Moving beyond the grounding issue. *Psychonomic Bulletin & Review*, 23(4), 991–1001. <https://doi.org/10.3758/s13423-015-0870-z>
- Lewis, J. W., Phinney, R. E., Brefczynski-Lewis, J. A., & DeYoe, E. A. (2006). Lefties get it “right” when hearing tool sounds. *Journal of Cognitive Neuroscience*, 18(8), 1314-1330. <https://doi.org/10.1162/jocn.2006.18.8.1314>
- Lewis, R., & Noppeney, U. (2010). Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. *Journal of Neuroscience*, 30(37), 12329-12339. <https://doi.org/10.1523/JNEUROSCI.5745-09.2010>
- Li, A. Y., Ladyka-Wojcik, N., Qazilbash, H., Golestani, A., Walther, D. B., Martin, C. B., & Barense, M. D. (2024). Experience transforms crossmodal object representations in the anterior temporal lobes. *eLife*, 13, e83382. <https://doi.org/10.7554/eLife.83382>
- Li, A. Y., Liang, J. C., Lee, A. C., & Barense, M. D. (2020). The validated circular shape space: Quantifying the visual similarity of shape. *Journal of Experimental Psychology: General*, 149(5), 949–966. <https://doi.org/10.1037/xge0000693>
- Li, J., & Deng, S. W. (2023). Facilitation and interference effects of the multisensory context on learning: a systematic review and meta-analysis. *Psychological Research*, 87(5), 1334-1352. <https://doi.org/10.1007/s00426-022-01733-4>
- Lieberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431 -461. <https://doi.org/10.1037/h0020279>
- Lieberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5), 358–368. <https://doi.org/10.1037/h0044417>

- Lim, S. J., Fiez, J. A., & Holt, L. L. (2019). Role of the striatum in incidental learning of sound categories. *Proceedings of the National Academy of Sciences*, 116(10), 4671-4680. <https://doi.org/10.1073/pnas.181199211>
- List, A., Iordanescu, L., Grabowecky, M., & Suzuki, S. (2014). Haptic guidance of overt visual attention. *Attention, Perception, & Psychophysics*, 76(8), 2221-2228. <https://doi.org/10.3758/s13414-014-0696-1>
- Liu, Q., Zhang, Y., Campos, J. L., Zhang, Q., & Sun, H. J. (2011). Neural mechanisms for the effect of prior knowledge on audiovisual integration. *Biological Psychology*, 87(2), 200-208. <https://doi.org/10.1016/j.biopsycho.2011.02.006>
- Liu, S. T., Ma, L., & Wang, X. (2019). Optimal features for auditory categorization. *Nature Communications*, 10(1), 1225. <https://doi.org/10.1038/s41467-019-09115-y>
- Liu, T., & Cooper, L. A. (2003). Explicit and implicit memory for rotating objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(4), 554-562. <https://doi.org/10.1037/0278-7393.29.4.554>
- Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(3), 732-753. <https://doi.org/10.1037//0278-7393.24.3.732>
- Lotto, A. J., Holt, L. L., & Kluender, K. R. (1999). Structure of phonetic categories produced by general learning mechanisms. *The Journal of the Acoustical Society of America*, 106(4_Supplement), 2247-2247. <https://doi.org/10.1121/1.427665>
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, 111(2), 309-332. <https://doi.org/10.1037/0033-295X.111.2.309>
- Lu, K., Xu, Y., Yin, P., Oxenham, A. J., Fritz, J. B., & Shamma, S. A. (2017). Temporal coherence structure rapidly shapes neuronal interactions. *Nature Communications*, 8(1), 13900. <https://doi.org/10.1038/ncomms13900>
- Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes*, 66(3), 309-332. <https://doi.org/10.1016/j.beproc.2004.03.011>
- Maddox, W. T., Ing, A. D., & Lauritzen, J. S. (2006). Stimulus modality interacts with category structure in perceptual category learning. *Perception & Psychophysics*, 68(7), 1176-1190. <https://doi.org/10.3758/BF03193719>

- Mädebach, A., Wöhner, S., Kieseler, M. L., & Jescheniak, J. D. (2017). Neighing, barking, and drumming horses-object related sounds help and hinder picture naming. *Journal of Experimental Psychology. Human Perception and Performance*, 43(9), 1629–1646. <https://doi.org/10.1037/xhp0000415>
- Mahns, D. A., Perkins, N. M., Sahai, V., Robinson, L., & Rowe, M. J. (2006). Vibrotactile frequency discrimination in human hairy skin. *Journal of Neurophysiology*, 95(3), 1442-1450. <https://doi.org/10.1152/jn.00483.2005>
- Mahon, B. Z. (2015). What is embodied about cognition?. *Language, Cognition and Neuroscience*, 30(4), 420-429. <https://doi.org/10.1080/23273798.2014.987791>
- Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology, Paris*, 102(1-3), 59–70. <https://doi.org/10.1016/j.jphysparis.2008.03.004>
- Mair, P., Groenen, P. J., & de Leeuw, J. (2022). More on multidimensional scaling and unfolding in R: smacof version 2. *Journal of Statistical Software*, 102 (10), 1-47. <https://doi.org/10.18637/jss.v102.i10>
- Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., ... & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, 92(18), 8135-8139. <https://doi.org/10.1073/pnas.92.18.8135>
- Malone, P. S., Eberhardt, S. P., Wimmer, K., Sprouse, C., Klein, R., Glomb, K., ... & Riesenhuber, M. (2019). Neural mechanisms of vibrotactile categorization. *Human Brain Mapping*, 40(10), 3078-3090. <https://doi.org/10.1002/hbm.24581>
- Manfredi, L. R., Saal, H. P., Brown, K. J., Zielinski, M. C., Dammann III, J. F., Polashock, V. S., & Bensmaia, S. J. (2014). Natural scenes in tactile texture. *Journal of Neurophysiology*, 111(9), 1792-1802. <https://doi.org/10.1152/jn.00680.2013>
- Marin-Campos, R., Dalmau, J., Compte, A., & Linares, D. (2021). StimuliApp: Psychophysical tests on mobile devices. *Behavior Research Methods*, 53(3), 1301-1307. <https://doi.org/10.3758/s13428-020-01491-4>
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press. <https://doi.org/10.7551/mitpress/9780262514620.001.0001>
- Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 200(1140), 269-294. <http://doi.org/10.1098/rspb.1978.0020>
- Martin, A. (2007). The representation of object concepts in the brain. *Annu. Rev. Psychol.*, 58(1), 25-45. <https://doi.org/10.1146/annurev.psych.57.102904.190143>

- Marucci, M., Di Flumeri, G., Borghini, G., Sciaraffa, N., Scandola, M., Pavone, E. F., ... & Aricò, P. (2021). The impact of multisensory integration and perceptual load in virtual reality settings on performance, workload and presence. *Scientific Reports*, 11(1), 4831. <https://doi.org/10.1038/s41598-021-84196-8>
- Masson, H. L., Bulthé, J., De Beeck, H. P. O., & Wallraven, C. (2016). Visual and Haptic Shape Processing in the Human Brain: Unisensory Processing, Multisensory Convergence, and Top-Down Influences. *Cerebral Cortex*, 26(8). <https://doi.org/10.1093/cercor/bhv170>
- MathWorks. (2023). *MATLAB* (Version 23.2.0.2380103 R2023b) [Computer software]. The MathWorks Inc.
- Matusz, P. J., Merkley, R., Faure, M., & Scerif, G. (2019). Expert attention: Attentional allocation depends on the differential development of multisensory number representations. *Cognition*, 186, 171–177. <https://doi.org/10.1016/j.cognition.2019.01.013>
- Matusz, P. J., Wallace, M. T., & Murray, M. M. (2017). A multisensory perspective on object memory. *Neuropsychologia*, 105, 243–252. <https://doi.org/10.1016/j.neuropsychologia.2017.04.008>
- McAdams, S. (1993). Recognition of sound sources and events. *Thinking in Sound: The Cognitive Psychology of Human Audition*, 146–198. <https://doi.org/10.1093/acprof:oso/9780198522577.003.0006>
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748. <https://doi.org/10.1038/264746a0>
- McMurray B. (2022). The myth of categorical perception. *The Journal of the Acoustical Society of America*, 152(6), 3819. <https://doi.org/10.1121/10.0016614>
- McMurray, B., & Chiu, S. (2020). Reinforced statistical learning of auditory categories. <https://doi.org/10.31234/osf.io/rzbfu>
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3), 207–238. <https://doi.org/10.1037/0033-295X.85.3.207>
- Mercier & Cappe, 2020
- Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, 19(2), 242–279. [https://doi.org/10.1016/0010-0285\(87\)90012-0](https://doi.org/10.1016/0010-0285(87)90012-0)
- Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science*, 221(4608), 389–391. <https://doi.org/10.1126/science.6867718>
- Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology*, 56(3), 640–662. <https://doi.org/10.1152/jn.1986.56.3.640>

- Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, 32, 89–115. <https://doi.org/10.1146/annurev.ps.32.020181.000513>
- Metzger, A., & Drewing, K. (2019). Memory influences haptic perception of softness. *Scientific Reports*, 9(1), 14383. <https://doi.org/10.1038/s41598-019-50835-4>
- Michel, C. (2021). Overcoming the modal/amodal dichotomy of concepts. *Phenomenology and the Cognitive Sciences*, 20(4), 655-677. <https://doi.org/10.1007/s11097-020-09678-y>
- Micheyl, C., Hanson, C., Demany, L., Shamma, S., & Oxenham, A. J. (2013). Auditory stream segregation for alternating and synchronous tones. *Journal of Experimental Psychology. Human Perception and Performance*, 39(6), 1568–1580. <https://doi.org/10.1037/a0032241>
- Miller, R. L., & Sheinberg, D. L. (2022). Evidence for independent processing of shape by vision and touch. *eNeuro*, 9(3). <https://doi.org/10.1523/ENEURO.0502-21.2022>
- Mishkin M. (1979). Analogous neural models for tactual and visual learning. *Neuro Psychologia*, 17, 139–151. [https://doi.org/10.1016/0028-3932\(79\)90005-8](https://doi.org/10.1016/0028-3932(79)90005-8)
- Mishkin, M., & Ungerleider, L. G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioural Brain Research*, 6(1), 57-77. [https://doi.org/10.1016/0166-4328\(82\)90081-X](https://doi.org/10.1016/0166-4328(82)90081-X)
- 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. <https://doi.org/10.1093/cercor/bhh007>
- Molholm, S., Sehatpour, P., Mehta, A. D., Shpaner, M., Gomez-Ramirez, M., Ortigue, S., Dyke, J. P., Schwartz, T. H., & Foxe, J. J. (2006). Audio-visual multisensory integration in superior parietal lobule revealed by human intracranial recordings. *Journal of Neurophysiology*, 96(2), 721–729. <https://doi.org/10.1152/jn.00285.2006>
- Montano-Murillo, R., Pittera, D., Frier, W., Georgiou, O., Obrist, M., & Cornelio, P. (2023). It Sounds Cool: Exploring Sonification of Mid-Air Haptic Textures Exploration on Texture Judgments, Body Perception, and Motor Behaviour. *IEEE Transactions on Haptics*, PP, <http://doi.org/10.1109/TOH.2023.3320492>
- Moran, Z. D., Bachman, P., Pham, P., Cho, S. H., Cannon, T. D., & Shams, L. (2013). Multisensory encoding improves auditory recognition. *Multisensory Research*, 26(6), 581–592. <https://doi.org/10.1163/22134808-00002436>
- Morgenstern, Y., Schmidt, F., & Fleming, R. W. (2019). One-shot categorization of novel object classes in humans. *Vision Research*, 165, 98-108. <https://doi.org/10.1016/j.visres.2019.09.005>

- Muender, T., Bonfert, M., Reinschluessel, A. V., Malaka, R., & Döring, T. (2022). Haptic fidelity framework: Defining the factors of realistic haptic feedback for virtual reality. In *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1-17). <https://doi.org/10.1145/3491102.3501953>
- Mulatti, C., Treccani, B., & Job, R. (2014). The role of the sound of objects in object identification: evidence from picture naming. *Frontiers in Psychology, 5*, 1139. <https://doi.org/10.3389/fpsyg.2014.01139>
- Muller, C. J., & MacLehose, R. F. (2014). Estimating predicted probabilities from logistic regression: different methods correspond to different target populations. *International Journal of Epidemiology, 43*(3), 962-970. <https://doi.org/10.1093/ije/dyu029>
- Müller, H. (2013). The real-exposure effect revisited—How purchase rates vary under pictorial vs. real item presentations when consumers are allowed to use their tactile sense. *International Journal of Research in Marketing, 30*(3), 304-307. <https://doi.org/10.1016/j.ijresmar.2013.03.001>
- Murphy, G. (2004). *The big book of concepts*. MIT press. Murray & Sperdin. <https://doi.org/10.7551/mitpress/1602.001.0001>
- Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 904-919. <https://doi.org/10.1037/0278-7393.20.4.904>
- Murphy, G., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review, 92* (3), pp. 289-316. <https://doi.org/10.1037/0033-295x.92.3.289>
- Murray, D. J., Ellis, R. R., Bandomir, C. A., & Ross, H. E. (1999). Charpentier (1891) on the size—weight illusion. *Perception & Psychophysics, 61*(8), 1681-1685. <https://doi.org/10.3758/BF03213127>
- Murray, M. M., & Sperdin, H. F. (2010). Single-trial multisensory learning and memory retrieval. *Multisensory Object Perception in the Primate Brain*, 191-208. https://doi.org/10.1007/978-1-4419-5615-6_11
- Murray, M. M., Camen, C., Andino, S. L. G., Bovet, P., & Clarke, S. (2006). Rapid brain discrimination of sounds of objects. *Journal of Neuroscience, 26*(4), 1293-1302. <https://doi.org/10.1523/JNEUROSCI.4511-05.2006>
- Murray, M. M., Eardley, A. F., Edginton, T., Oyekan, R., Smyth, E., & Matusz, P. J. (2018). Sensory dominance and multisensory integration as screening tools in aging. *Scientific Reports, 8*(1), 8901. <https://doi.org/10.1038/s41598-018-27288-2>

- Murray, M. M., Foxe, J. J., & Wylie, G. R. (2005). The brain uses single-trial multisensory memories to discriminate without awareness. *Neuroimage*, 27(2), 473-478. <https://doi.org/10.1016/j.neuroimage.2005.04.016>
- Murray, M. M., Lewkowicz, D. J., Amedi, A., & Wallace, M. T. (2016). Multisensory processes: a balancing act across the lifespan. *Trends in Neurosciences*, 39(8), 567-579. [https://doi.org/HYPERLINK](https://doi.org/HYPERLINK\) <https://doi.org/10.1016/j.tins.2016.05.003>
- Murray, M. M., Michel, C. M., Grave de Peralta, R., Ortigue, S., Brunet, D., Gonzalez Andino, S., & Schnider, A. (2004). Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging. *NeuroImage*, 21(1), 125–135. <https://doi.org/10.1016/j.neuroimage.2003.09.035>
- Murray, M. M., Molholm, S., Michel, C. M., Heslenfeld, D. J., Ritter, W., Javitt, D. C., ... & Foxe, J. J. (2005). Grabbing your ear: rapid auditory–somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment. *Cerebral Cortex*, 15(7), 963-974. <https://doi.org/10.1093/cercor/bhh197>
- Murray, S. O., Schrater, P., & Kersten, D. (2004). Perceptual grouping and the interactions between visual cortical areas. *Neural Networks*, 17(5-6), 695-705. <https://doi.org/10.1016/j.neunet.2004.03.010>
- Myers, E. B., Blumstein, S. E., Walsh, E., & Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychological science*, 20(7), 895-903. <https://doi.org/10.1111/j.1467-9280.2009.02380>
- Naci, L., Taylor, K. I., Cusack, R., & Tyler, L. K. (2012). Are the senses enough for sense? Early high-level feedback shapes our comprehension of multisensory objects. *Frontiers in Integrative Neuroscience*, 6, 82. <https://doi.org/10.3389/fnint.2012.00082>
- Nardini, M., Bedford, R., & Mareschal, D. (2010). Fusion of visual cues is not mandatory in children. *Proceedings of the National Academy of Sciences*, 107(39), 17041-17046. <https://doi.org/10.1073/pnas.1001699107>
- Neumann, P. G. (1974). An attribute frequency model for the abstraction of prototypes. *Memory & Cognition*, 2(2), 241-248. <https://doi.org/10.3758/BF03208990>
- Newell, F. N. (1998). Stimulus context and view dependence in object recognition. *Perception*, 27(1), 47-68. <https://doi.org/10.1068/p270047>
- Newell, F. N. (2004). Cross-modal object recognition. *The Handbook of Multisensory Processes*, 123-139. Review. <https://doi.org/10.7551/mitpress/3422.003.0011>
- Newell, F. N., & Bülthoff, H. H. (2002). Categorical perception of familiar objects. *Cognition*, 85(2), 113–143. [https://doi.org/10.1016/s0010-0277\(02\)00104-x](https://doi.org/10.1016/s0010-0277(02)00104-x)

- Newell, F. N., Ernst, M. O., Tjan, B. S., & Bülthoff, H. H. (2001). Viewpoint dependence in visual and haptic object recognition. *Psychological Science*, 12(1), 37-42. <https://doi.org/10.1111/1467-9280.00307>
- Newell, F. N., McKenna, E., Seveso, M. A., Devine, I., Alahmad, F., Hirst, R. J., & O'Dowd, A. (2023). Multisensory perception constrains the formation of object categories: A review of evidence from sensory-driven and predictive processes on categorical decisions. *Philosophical Transactions of the Royal Society B*, 378(1886), 20220342. <https://doi.org/10.1098/rstb.2022.0342>
- Newell, F. N., Wallraven, C., & Huber, S. (2004). The role of characteristic motion in object categorization. *Journal of Vision*, 4(2), 5-5. <https://doi.org/10.1167/4.2.5>
- Newman, M. A. R. K., Gatersleben, B., Wyles, K. J., & Ratcliffe, E. (2022). The use of virtual reality in environment experiences and the importance of realism. *Journal of Environmental Psychology*, 79, 101733. <https://doi.org/10.1016/j.jenvp.2021.101733>
- Noppeney, U. (2021). Perceptual inference, learning, and attention in a multisensory world. *Annual Review of Neuroscience*, 44(1), 449-473. <https://doi.org/10.1146/annurev-neuro-100120-085519>
- Nordmark, J. O. (1978). Frequency and periodicity analysis. *Handbook of Perception*, 243-282. <https://doi.org/10.1016/B978-0-12-161904-6.50014-2>
- Norman, J. F., & Todd, J. T. (1993). The perceptual analysis of structure from motion for rotating objects undergoing affine stretching transformations. *Perception & Psychophysics*, 53(3), 279-291. <https://doi.org/10.3758/BF03205183>
- Norman, J. F., Clayton, A. M., Norman, H. F., & Crabtree, C. E. (2008). Learning to perceive differences in solid shape through vision and touch. *Perception*, 37(2), 185–196. <https://doi.org/10.1068/p5679>
- Norman, J. F., Norman, H. F., Clayton, A. M., Lianekhammy, J., & Zielke, G. (2004). The visual and haptic perception of natural object shape. *Perception & psychophysics*, 66(2), 342-351. <https://doi.org/10.3758/bf03194883>
- Norman, J. F., Todd, J. T., & Orban, G. A. (2004). Perception of three-dimensional shape from specular highlights, deformations of shading, and other types of visual information. *Psychological Science*, 15(8), 565-570. <https://doi.org/10.1111/j.0956-7976.2004.00720.x>
- Nosofsky R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology. General*, 115(1), 39–61. <https://doi.org/10.1037//0096-3445.115.1.39>
- Nosofsky, R. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, 43, 25-53. <https://doi.org/10.1146/annurev.ps.43.020192.000325>

- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1), 39-61. <https://doi.org/10.1037//0096-3445.115.1.39>
- Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(1), 87-108. <https://doi.org/10.1037/0278-7393.13.1.87>
- Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, 104(2), 266-300. <https://doi.org/10.1037/0033-295x.104.2.266>
- Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(2), 282-304. <http://dx.doi.org/10.1037/0278-7393.15.2.282>
- Notman, L. A., Sowden, P. T., & Özgen, E. (2005). The nature of learned categorical perception effects: a psychophysical approach. *Cognition*, 95(2), B1-B14. <https://doi.org/10.1016/j.cognition.2004.07.002>
- Nuiten, S. A., Fahrenfort, J. J., & Bekinschtein, T. A. (2021). Preserved sensory processing but hampered conflict detection when stimulus input is task-irrelevant. *eLife*, 10, e64431. <https://doi.org/10.7554/eLife.64431>
- Nyberg, L., Habib, R., McIntosh, A. R., & Tulving, E. (2000). Reactivation of encoding-related brain activity during memory retrieval. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20), 11120–11124. <https://doi.org/10.1073/pnas.97.20.11120>
- O'Bryan, S. R., Jung, S., Mohan, A. J., & Scolari, M. (2024). Category learning selectively enhances representations of boundary-adjacent exemplars in early visual cortex. *Journal of Neuroscience*, 44(3). <https://doi.org/10.1523/JNEUROSCI.1039-23.2023>
- O'Callaghan, C. (2008). Object perception: Vision and audition. *Philosophy Compass*, 3(4), 803-829. <https://doi.org/10.1111/j.1747-9991.2008.00145.x>
- Obasih, C. O., Luthra, S., Dick, F., & Holt, L. L. (2023). Auditory category learning is robust across training regimes. *Cognition*, 237, 105467. <https://doi.org/10.1016/j.cognition.2023.105467>
- O'Callaghan, G., O'Dowd, A., Simões-Franklin, C., Stapleton, J., & Newell, F. N. (2018). Tactile-to-Visual Cross-Modal Transfer of Texture Categorisation Following Training: An fMRI Study. *Frontiers in Integrative Neuroscience*, 12, 24. <https://doi.org/10.3389/fnint.2018.00024>
- O'Dowd, A., Hirst, R. J., Seveso, M. A., McKenna, E. M., & Newell, F. N. (2025). Generalisation to novel exemplars of learned shape categories based on visual and auditory spatial cues does not benefit from multisensory information. *Psychonomic Bulletin & Review*, 32(1), 417–429. <https://doi.org/10.3758/s13423-024-02548-7>

- Ogawa, N., Narumi, T., & Hirose, M. (2019, March). Virtual hand realism affects object size perception in body-based scaling. In *2019 IEEE conference on virtual reality and 3D user interfaces (VR)*, 519-528. <https://doi.org/10.1109/VR.2019.8798040> .
- Op de Beeck, H., Wagemans, J., & Vogels, R. (2003). The effect of category learning on the representation of shape: dimensions can be biased but not differentiated. *Journal of Experimental Psychology: General*, 132(4), 491–511. <https://doi.org/10.1037/0096-3445.132.4.491>
- O'Regan, J. K., & Noë, A. (2001). What it is like to see: A sensorimotor theory of perceptual experience. *Synthese*, 129(1), 79-103. <https://doi.org/10.1023/A:1012699224677>
- O'Regan, J., Rensink, R. & Clark, J. (1999). Change-blindness as a result of 'mudsplashes'. *Nature* 398(6722), 34-34. <https://doi.org/10.1038/17953>
- Oruç, I., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Research*, 43(23), 2451–2468. [https://doi.org/10.1016/s0042-6989\(03\)00435-8](https://doi.org/10.1016/s0042-6989(03)00435-8)
- Özgen, E., & Davies, I. R. (2002). Acquisition of categorical color perception: a perceptual learning approach to the linguistic relativity hypothesis. *Journal of Experimental Psychology: General*, 131(4), 477-493. <https://doi.org/10.1037/0096-3445.131.4.477>
- Palacios-Navarro, G., & Hogan, N. (2021). Head-mounted display-based therapies for adults post-stroke: a systematic review and meta-analysis. *Sensors*, 21(4), 1111. <https://doi.org/10.3390/s21041111>
- Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. *Nature Communications*, 7, 11543. <https://doi.org/10.1038/ncomms11543>
- Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in Human Neuroscience*, 9, 660 <https://doi.org/10.3389/fnhum.2015.00660>
- Pascual-Leone, A., & Hamilton, R. (2001). The metamodal organization of the brain. *Progress in Brain Research*, 134, 427–445. [https://doi.org/10.1016/S0079-6123\(01\)34028-1](https://doi.org/10.1016/S0079-6123(01)34028-1)
- Pastel, S., Bürger, D., Chen, C. H., Petri, K., & Witte, K. (2022). Comparison of spatial orientation skill between real and virtual environment. *Virtual Reality*, 26(1), 91-104. <https://doi.org/10.1007/s10055-021-00539-w>
- Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. <https://doi.org/10.48550/arXiv.1503.06462>

- Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8(12), 976-987. <https://doi.org/10.1038/nrn2277>
- Pecher, D., & Zeelenberg, R. (2022). Does multisensory study benefit memory for pictures and sounds?. *Cognition*, 226, 105181. <https://doi.org/10.1016/j.cognition.2022.105181>
- Peelen, M. V., & Downing, P. E. (2017). Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia*, 105, 177-183. <https://doi.org/10.1016/j.neuropsychologia.2017.03.033>
- Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4), 1643-1662. <https://doi.org/10.3758/s13428-021-01694-3>.
- Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Peirce, J., Hirst, R., & MacAskill, M. (2022). *Building experiments in PsychoPy*. Sage.
- Peissig, J. J., & Tarr, M. J. (2007). Visual object recognition: do we know more now than we did 20 years ago?. *Annual Review of Psychology*, 58, 75–96. <https://doi.org/10.1146/annurev.psych.58.102904.190114>
- Pérez-Bellido, A., Anne Barnes, K., Crommett, L. E., & Yau, J. M. (2018). Auditory frequency representations in human somatosensory cortex. *Cerebral Cortex*, 28(11), 3908-3921. <https://doi.org/10.1093/cercor/bhx255>
- Pérez-Bellido, A., Pappal, R. D., & Yau, J. M. (2018). Touch engages visual spatial contextual processing. *Scientific Reports*, 8(1), 16637. <https://doi.org/10.1038/s41598-018-34810-z>
- Pérez-Gay Juárez, F., Sicotte, T., Thériault, C., & Harnad, S. (2019). Category learning can alter perception and its neural correlates. *PLoS One*, 14(12), e0226000. <https://doi.org/10.1371/journal.pone.0226000>
- Pérez-Gay Juárez, F., Thériault, C., Gregory, M., Rivas, D., Sabri, H., & Harnad, S. (2017). How and why does category learning cause categorical perception?. *International Journal of Comparative Psychology*, 30-62. <https://doi.org/10.46867/ijcp.2017.30.01.01>
- Perini, F., Powell, T., Watt, S. J., & Downing, P. E. (2020). Neural representations of haptic object size in the human brain revealed by multivoxel fMRI patterns. *Journal of Neurophysiology*, 124(1), 218-231. <https://doi.org/10.1152/jn.00160.2020>

- Petkov, G., & Petrova, Y. (2019). Relation-based categorization and category learning as a result from structural alignment. The RoleMap model. *Frontiers in Psychology*, 10, 563. <https://doi.org/10.3389/fpsyg.2019.00563>
- Phillips, F., Egan, E. J. L., & Perry, B. N. (2009). Perceptual equivalence between vision and touch is complexity dependent. *Acta Psychologica*, 132(3), 259-266. <https://doi.org/10.1016/j.actpsy.2009.07.010>
- Pietrini, P., Furey, M. L., Ricciardi, E., Gobbin, M. I., Wu, W. H. C., Cohen, L., ... & Haxby, J. V. (2004). Beyond sensory images: Object-based representation in the human ventral pathway. *Proceedings of the National Academy of Sciences*, 101(15), 5658-5663. <https://doi.org/10.1073/pnas.0400707101>
- Pinotsis, D. A., Siegel, M., & Miller, E. K. (2019). Sensory Processing and Categorization in Cortical and Deep Neural Networks. *NeuroImage*, 202, 116118. <https://doi.org/10.1016/j.neuroimage.2019.116118>
- Poggio, T., & Girosi, F. (2002). Networks for approximation and learning. *Proceedings of the IEEE*, 78(9), 1481-1497. <https://doi.org/10.1109/5.58326>
- Pomper, U., Brincker, J., Harwood, J., Prihodko, I., & Senkowski, D. (2014). Taking a call is facilitated by the multisensory processing of smartphone vibrations, sounds, and flashes. *PloS One*, 9(8), e103238. <https://doi.org/10.1371/journal.pone.0103238>
- Posner, M. I., & Keele, S. W. (1968). On the Genesis of Abstract Ideas. *Journal of Experimental Psychology*, 77(3), 353-363. <https://doi.org/10.1037/h0025953>
- Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, 83(2, Pt.1), 304-308. <https://doi.org/10.1037/h0028558>
- 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), 157-171. <https://doi.org/10.1037/0033-295X.83.2.157>
- Potter M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology. Human Learning and Memory*, 2(5), 509-522. <https://doi.org/10.1037/0278-7393.2.5.509>
- Potter M. C. (1993). Very short-term conceptual memory. *Memory & Cognition*, 21(2), 156-161. <https://doi.org/10.3758/bf03202727>
- Power, R. P., & Graham, A. (1976). Dominance of touch by vision: generalization of the hypothesis to a tactually experienced population. *Perception*, 5(2), 161-166. <https://doi.org/10.1068/p050161>

- Pulvermüller, F. (2013). How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, 17(9), 458-470. <https://doi.org/10.1016/j.tics.2013.06.004>
- Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, 3(1), 111-132. <https://doi.org/10.1017/S0140525X00002053>
- Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge, MA:MIT Press. <https://doi.org/10.1017/S0012217300049854>
- R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rauss, K., & Pourtois, G. (2013). What is bottom-up and what is top-down in predictive coding?. *Frontiers in Psychology*, 4, 276. <https://doi.org/10.3389/fpsyg.2013.00276>
- Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3(3), 382-407. [https://doi.org/10.1016/0010-0285\(72\)90014-X](https://doi.org/10.1016/0010-0285(72)90014-X)
- Rezaeizadeh, M., & Shamma, S. (2021). Binding the acoustic features of an auditory source through temporal coherence. *Cerebral Cortex Communications*, 2(4), tgab060. <https://doi.org/10.1093/texcom/tgab060>
- Rice, N. J., Tunik, E., & Grafton, S. T. (2006). The anterior intraparietal sulcus mediates grasp execution, independent of requirement to update: new insights from transcranial magnetic stimulation. *Journal of Neuroscience*, 26(31), 8176-8182. <https://doi.org/10.1523/JNEUROSCI.1641-06.2006>
- Riesenhuber, M. (2009). Object Categorization in Man, Monkey, and Machine: Some Answers and Some Open Questions. *Object Categorization: Computer and Human Vision Perspectives*, 216–240. <https://doi.org/10.1017/CBO9780511635465.013>
- Riva, G., Wiederhold, B. K., & Mantovani, F. (2019). Neuroscience of virtual reality: from virtual exposure to embodied medicine. *Cyberpsychology, Behavior, and Social Networking*, 22(1), 82-96. <https://doi.org/10.1089/cyber.2017.29099.gri>
- Roark, C. L. (2024). Perceptual category learning results in modality-specific representations. In *Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 46)*. <https://escholarship.org/uc/item/236389bq>
- Roark, C. L., & Chandrasekaran, B. (2023). Stable, flexible, common, and distinct behaviors support rule-based and information-integration category learning. *npj Science of Learning*, 8(1), 14. <https://doi.org/10.1038/s41539-023-00163-0>

- Roark, C. L., & Holt, L. L. (2022). Long-term priors constrain category learning in the context of short-term statistical regularities. *Psychonomic Bulletin & Review*, 29(5), 1925–1937. <https://doi.org/10.3758/s13423-022-02114-z>
- Roark, C. L., Lescht, E., Hampton Wray, A., & Chandrasekaran, B. (2023). Auditory and visual category learning in children and adults. *Developmental Psychology*, 59(5), 963–975. <https://doi.org/10.1037/dev0001525>
- Roark, C. L., Paulon, G., Sarkar, A., & Chandrasekaran, B. (2021). Comparing perceptual category learning across modalities in the same individuals. *Psychonomic Bulletin & Review*, 28(3), 898–909. <https://doi.org/10.3758/s13423-021-01878-0>
- Roberson, D., Hanley, J. R., & Pak, H. (2009). Thresholds for color discrimination in English and Korean speakers. *Cognition*, 112(3), 482–487. <https://doi.org/10.1016/j.cognition.2009.06.008>
- Robert, S., Ungerleider, L. G., & Vaziri-Pashkam, M. (2023). Disentangling Object Category Representations Driven by Dynamic and Static Visual Input. *The Journal of Neuroscience: the official journal of the Society for Neuroscience*, 43(4), 621–634. <https://doi.org/10.1523/JNEUROSCI.0371-22.2022>
- Rock, I., & Victor, J. (1964). Vision and touch: An experimentally created conflict between the two senses. *Science*, 143(3606), 594–596. <https://doi.org/10.1126/science.143.3606.594>
- Röder, B., Stock, O., Bien, S., Neville, H., & Rösler, F. (2002). Speech processing activates visual cortex in congenitally blind humans. *European Journal of Neuroscience*, 16(5), 930–936. <https://doi.org/10.1046/j.1460-9568.2002.02147.x>
- Rohe, T., & Noppeney, U. (2015). Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLoS biology*, 13(2), e1002073. <https://doi.org/10.1371/journal.pbio.1002073>
- Rohe, T., & Noppeney, U. (2015). Sensory reliability shapes perceptual inference via two mechanisms. *Journal of Vision*, 15(5), 22. <https://doi.org/10.1167/15.5.22>
- Rohe, T., Ehrlis, A. C., & Noppeney, U. (2019). The neural dynamics of hierarchical Bayesian causal inference in multisensory perception. *Nature Communications*, 10(1), 1907. <https://doi.org/10.1038/s41467-019-09664-2>
- Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3), 192–233. <https://doi.org/10.1037/0096-3445.104.3.192>
- Rosch, E. (1978). Principles of categorization. In *Cognition and Categorization* (pp. 27–48). Routledge.
- Rosch, E. (1999). Reclaiming concepts. *Journal of Consciousness Studies*, 6(11–12), 61–77.

- Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, 4(3), 328-350. [https://doi.org/10.1016/0010-0285\(73\)90017-0](https://doi.org/10.1016/0010-0285(73)90017-0)
- Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4), 573-605. [https://doi.org/10.1016/0010-0285\(75\)90024-9](https://doi.org/10.1016/0010-0285(75)90024-9)
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3), 382-439. [https://doi.org/10.1016/0010-0285\(76\)90013-X](https://doi.org/10.1016/0010-0285(76)90013-X)
- Roth, D., & Latoschik, M. E. (2020). Construction of the virtual embodiment questionnaire (veq). *IEEE Transactions on Visualization and Computer Graphics*, 26(12), 3546-3556. <https://doi.org/10.1109/TVCG.2020.3023603>
- Roudaia, E., Calabro, F. J., Vaina, L. M. and Newell, F. N. (2018). Aging impairs audio-visual facilitation of object motion within self-motion, *Multisens. Res.* 31, 251–272. <https://doi.org/10.1163/22134808-00002600>
- Rowland, B. A., Quessy, S., Stanford, T. R., & Stein, B. E. (2007). Multisensory integration shortens physiological response latencies. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, 27(22), 5879–5884. <https://doi.org/10.1523/JNEUROSCI.4986-06.2007>
- Rubinstein, G. (2008). On sounds emitted by inanimate objects in Russian. *The Slavic and East European Journal*, 52(4), 561–588. <https://www.jstor.org/stable/40651272>
- Russell, B. (1918). The relation of sense data to physics. In *Mysticism and logic*, 139-170.
- Ryan, C. P., Bettelani, G. C., Ciotti, S., Parise, C., Moscatelli, A., & Bianchi, M. (2021). The interaction between motion and texture in the sense of touch. *Journal of Neurophysiology*, 126(4), 1375–1390. <https://doi.org/10.1152/jn.00583.2020>
- Sanabria, D., Correa, Á., Lupiáñez, J., & Spence, C. (2004). Bouncing or streaming? Exploring the influence of auditory cues on the interpretation of ambiguous visual motion. *Experimental Brain Research*, 157(4), 537-541. <https://doi.org/10.1007/s00221-004-1993-z>
- Sanchez-Vives, M. V., & Slater, M. (2005). From presence to consciousness through virtual reality. *Nature Reviews Neuroscience*, 6(4), 332–339. <https://doi.org/10.1038/nrn1651>
- Sandhu, R., & Dyson, B. J. (2012). Re-evaluating visual and auditory dominance through modality switching costs and congruency analyses. *Acta Psychologica*, 140(2), 111-118. <https://doi.org/10.1016/j.actpsy.2012.04.003>

- Sathian, K. (2016). Analysis of haptic information in the cerebral cortex. *Journal of Neurophysiology*, 116(4), 1795-1806. <https://doi.org/10.1152/jn.00546.2015>
- Sathian, K., & Lacey, S. (2022). Cross-modal interactions of the tactile system. *Current Directions in Psychological Science*, 31(5), 411-418. <https://doi.org/10.1177/09637214221101877>
- Sathian, K., Lacey, S., Stilla, R., Gibson, G., Deshpande, G., Hu, X., et al. (2011). Dual pathways for somatosensory information processing. *Neuroimage* 57, 462–475. <https://doi.org/10.1016/j.neuroimage.2011.05.001>
- Schendan, H. E., & Ganis, G. (2015). Top-down modulation of visual processing and knowledge after 250 ms supports object constancy of category decisions. *Frontiers in Psychology*, 6, 1289. <https://doi.org/10.3389/fpsyg.2015.01289>
- Schmied-Kowarzik, R., Reisewitz, P., Kaschub, L., Rodeck, R., & Wende, G. (2023). An approach for visual realism complexity classification of 3d models in virtual and augmented reality. In *International Conference on Human-Computer Interaction*, 192-202. https://doi.org/10.1007/978-3-031-35634-6_14
- Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, 80(1-2), 1-46. [https://doi.org/10.1016/S0010-0277\(00\)00152-9](https://doi.org/10.1016/S0010-0277(00)00152-9)
- Scholl, B. J. (2007). Object persistence in philosophy and psychology. *Mind & Language*, 22(5), 563-591. <https://doi.org/10.1111/j.1468-0017.2007.00321.x>
- Scholl, B. J. (2009). What have we learned about attention from multiple-object tracking (and vice versa). *Computation, Cognition, and Pylyshyn*, 49–77. Boston Review.
- Schroeder, C. E., & Foxe, J. (2005). Multisensory contributions to low-level, 'unisensory' processing. *Current Opinion in Neurobiology*, 15(4), 454-458. <https://doi.org/10.1016/j.conb.2005.06.008>
- Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments*, 10(3), 266-281. <https://doi.org/10.1162/105474601300343603>
- Schwarzer, G., Küfer, I., & Wilkening, F. (1999). Learning categories by touch: On the development of holistic and analytic processing. *Memory & Cognition*, 27(5), 868-877. <https://doi.org/10.3758/BF03198539>
- Schyns, P. G. (1998). Diagnostic recognition: task constraints, object information, and their interactions. *Cognition*, 67(1-2), 147-179. [https://doi.org/10.1016/S0010-0277\(98\)00016-X](https://doi.org/10.1016/S0010-0277(98)00016-X)

- Sciutti, A., Damonte, F., Alloisio, M., & Sandini, G. (2019). Visuo-haptic exploration for multimodal memory. *Frontiers in Integrative Neuroscience*, 13, 15. <https://doi.org/10.3389/fnint.2019.00015>
- Seger, C. A., & Miller, E. K. (2010). Category learning in the brain. *Annual review of neuroscience*, 33(1), 203-219. <https://doi.org/10.1146/annurev.neuro.051508.135546>
- Seitz, A. R., Nanez, J. E., Sr, Holloway, S. R., & Watanabe, T. (2006). Perceptual learning of motion leads to faster flicker perception. *PloS One*, 1(1), e28. <https://doi.org/10.1371/journal.pone.0000028>
- Setti, A., & Newell, F. N. (2010). The effect of body and part-based motion on the recognition of unfamiliar objects. *Visual Cognition*, 18(3), 456-480. <https://doi.org/10.1080/13506280902830561>
- Sewell, D. K., & Lewandowsky, S. (2011). Restructuring partitioned knowledge: The role of recoordination in category learning. *Cognitive Psychology*, 62(2), 81-122. <https://doi.org/10.1016/j.cogpsych.2010.09.003>
- Shamma, S. A., & Micheyl, C. (2010). Behind the scenes of auditory perception. *Current Opinion in Neurobiology*, 20(3), 361-366. <https://doi.org/10.1016/j.conb.2010.03.009>
- Shams, L., & Beierholm, U. (2022). Bayesian causal inference: A unifying neuroscience theory. *Neuroscience & Biobehavioral Reviews*, 137, 104619. <https://doi.org/10.1016/j.neubiorev.2022.104619>
- Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences*, 14(9), 425-432. <https://doi.org/10.1016/j.tics.2010.07.001>
- Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, 12(11), 411-417. <https://doi.org/10.1016/j.tics.2008.07.006>
- Shams, L., Ma, W. J., & Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *Neuroreport*, 16(17), 1923-1927. <https://doi.org/10.1097/01.wnr.0000187634.68504.bb>
- Shao, Y., Hayward, V., & Visell, Y. (2016). Spatial patterns of cutaneous vibration during whole-hand haptic interactions. *Proceedings of the National Academy of Sciences*, 113(15), 4188-4193. <https://doi.org/10.1073/pnas.1520866113>
- Shatek, S. M., Robinson, A. K., Grootswagers, T., & Carlson, T. A. (2022). Capacity for movement is an organisational principle in object representations. *NeuroImage*, 261, 119517. <https://doi.org/10.1016/j.neuroimage.2022.119517>
- Shepard R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820), 1317-1323. <https://doi.org/10.1126/science.3629243>

- Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1(1), 54-87. [https://doi.org/10.1016/0022-2496\(64\)90017-3](https://doi.org/10.1016/0022-2496(64)90017-3)
- Simanova, I., Francken, J. C., de Lange, F. P., & Bekkering, H. (2016). Linguistic priors shape categorical perception. *Language, Cognition and Neuroscience*, 31(1), 159-165. <https://doi.org/10.1080/23273798.2015.1072638>
- Simões-Franklin, C., Whitaker, T. A., & Newell, F. N. (2011). Active and passive touch differentially activate somatosensory cortex in texture perception. *Human Brain Mapping*, 32(7), 1067-1080. <https://doi.org/10.1002/hbm.21091>
- Siqueira Rodrigues, L., Schmidt, T. T., Israel, J. H., Zachow, S., Nyakatura, J., & Kosch, T. (2024). Comparing the Effects of Visual, Haptic, and Visuohaptic Encoding on Memory Retention of Digital Objects in Virtual Reality. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction*, 1-13. <https://doi.org/10.1145/3679318.3685349>
- Slater, M., Pérez Marcos, D., Ehrsson, H., & Sanchez-Vives, M. V. (2009). Inducing illusory ownership of a virtual body. *Frontiers in neuroscience*, 3, 214- 220. <https://doi.org/10.3389/neuro.01.029.2009>
- Slater, M., Pérez Marcos, D., Ehrsson, H., & Sanchez-Vives, M. V. (2008). Towards a digital body: the virtual arm illusion. *Frontiers in Human Neuroscience*, 2, 181. <https://doi.org/10.3389/neuro.09.006.2008>
- Smith, F. W., & Goodale, M. A. (2015). Decoding visual object categories in early somatosensory cortex. *Cerebral Cortex*, 25(4), 1020-1031. <https://doi.org/10.1093/cercor/bht292>
- Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4), 800–811. <https://doi.org/10.1037/0278-7393.28.4.800>
- Smith, J. D., Johnston, J. J., Musgrave, R. D., Zakrzewski, A. C., Boomer, J., Church, B. A., & Ashby, F. G. (2014). Cross-modal information integration in category learning. *Attention, Perception, & Psychophysics*, 76, 1473-1484. <https://doi.org/10.3758/s13414-014-0659-6>
- Smith, L. B., & Heise, D. (1992). Perceptual similarity and conceptual structure. In *Advances in Psychology* (93), 233-272. [https://doi.org/10.1016/S0166-4115\(08\)61009-2](https://doi.org/10.1016/S0166-4115(08)61009-2)
- Snow, J. C., Strother, L., & Humphreys, G. W. (2014). Haptic shape processing in visual cortex. *Journal of Cognitive Neuroscience*, 26(5), 1154-1167. https://doi.org/10.1162/jocn_a_00548
- Solso, R. L., & McCarthy, J. E. (1981). Prototype formation in a cognitive task. *British Journal of Psychology*, 72(4), 499–503. <https://doi.org/10.1111/j.2044-8295.1981.tb01779.x>

- Specht, J., Stegmann, B., Gross, H., & Krakow, K. (2023). Cognitive training with head-mounted display virtual reality in neurorehabilitation: Pilot randomized controlled trial. *JMIR Serious Games*, 11, e45816. <https://doi.org/10.2196/45816>
- Spence, C. (2007). Audiovisual multisensory integration. *Acoustical science and technology*, 28(2), 61-70. <https://doi.org/10.1250/ast.28.61>
- Spence, C., & Di Stefano, N. (2024). What, if anything, can be considered an amodal sensory dimension?. *Psychonomic Bulletin & Review*, 31(5), 1915-1933. <https://doi.org/10.3758/s13423-023-02447-3>
- Spence, C., & Driver, J. (2004). *Crossmodal space and crossmodal attention*. Oxford University Press.
- Sperdin, H. F., Cappe, C., Foxe, J. J., and Murray, M. M. (2009). Early, low-level auditory-somatosensory multisensory interactions impact reaction time speed. *Front. Integr. Neurosci.* 3:2 <https://doi.org/10.3389/neuro.07.002.2009>
- Sra, M., & Schmandt, C. (2015). Metaspace ii: Object and full-body tracking for interaction and navigation in social vr. *arXiv preprint arXiv:1512.02922*. <https://doi.org/10.48550/arXiv.1512.02922>
- Starke, J., Ball, F., Heinze, H. J., & Noesselt, T. (2020). The spatio-temporal profile of multisensory integration. *European Journal of Neuroscience*, 51(5), 1210-1223. <https://doi.org/10.1111/ejn.13753>
- Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. MIT press.
- Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9(4), 255-266. <https://doi.org/10.1038/nrn2331>
- Stein, B. E., Stanford, T. R., & Rowland, B. A. (2020). Multisensory integration and the society for neuroscience: Then and now. *Journal of Neuroscience*, 40(1), 3-11. <https://doi.org/10.1523/JNEUROSCI.0737-19.2019>
- Stevenson, R. A., & Wallace, M. T. (2013). Multisensory temporal integration: task and stimulus dependencies. *Experimental Brain Research*, 227(2), 249–261. <https://doi.org/10.1007/s00221-013-3507-3>
- Stevenson, R. A., Bushmakina, M., Kim, S., Wallace, M. T., Puce, A., & James, T. W. (2012). Inverse effectiveness and multisensory interactions in visual event-related potentials with audiovisual speech. *Brain Topography*, 25(3), 308–326. <https://doi.org/10.1007/s10548-012-0220-7>
- Stilla, R., & Sathian, K. (2008). Selective visuo-haptic processing of shape and texture. *Human Brain Mapping*, 29(10), 1123–1138. <https://doi.org/10.1002/hbm.20457>

- Stone, J. V. (1998). Object recognition using spatiotemporal signatures. *Vision Research*, 38(7), 947-951. [https://doi.org/10.1016/S0042-6989\(97\)00301-5](https://doi.org/10.1016/S0042-6989(97)00301-5)
- Stone, J. V. (1999). Object recognition: View-specificity and motion-specificity. *Vision Research*, 39(24), 4032-4044. [https://doi.org/10.1016/S0042-6989\(99\)00123-6](https://doi.org/10.1016/S0042-6989(99)00123-6)
- Störmer, V. S., McDonald, J. J., & Hillyard, S. A. (2009). Cross-modal cueing of attention alters appearance and early cortical processing of visual stimuli. *Proceedings of the National Academy of Sciences of the United States of America*, 106(52), 22456–22461. <https://doi.org/10.1073/pnas.0907573106>
- Su, H., Sun, M., Fei-Fei, L., & Savarese, S. (2009). Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In *2009 IEEE 12th International Conference on Computer Vision* (pp. 213-220). IEEE. <https://doi.org/10.1109/ICCV.2009.5459168>
- Suied, C., & Viaud-Delmon, I. (2009). Auditory-visual object recognition time suggests specific processing for animal sounds. *PloS One*, 4(4), e5256. <https://doi.org/10.1371/journal.pone.0005256>
- Sumser, A., Isaias-Camacho, E., Mease, R. A., & Groh, A. (2024). Differential representation of active and passive touch in mouse somatosensory thalamus. *bioRxiv*, 2024-07. <https://doi.org/10.1101/2024.07.16.603697>
- Sun, X., Yao, L., Fu, Q., & Fu, X. (2023). Multisensory transfer effects in implicit and explicit category learning. *Psychological Research*, 87(5), 1353-1369. <https://doi.org/10.1007/s00426-022-01754-z>
- Tajadura-Jimenez, A., Kitagawa, N., Valjamae, A., Zampini, M., Murray, M. M., and Spence, C. (2009). Auditory-somatosensory multisensory interactions are spatially modulated by stimulated body surface and acoustic spectra. *Neuropsychologia* 47, 195–203. <https://doi.org/10.1016/j.neuropsychologia.2008.07.025>
- Takahashi, C., & Watt, S. J. (2017). Optimal visual–haptic integration with articulated tools. *Experimental Brain Research*, 235(5), 1361-1373. <https://doi.org/10.1007/s00221-017-4896-5>
- Takeichi, H., Taniguchi, K., & Shigemasa, H. (2023). Visual and haptic cues in processing occlusion. *Frontiers in Psychology*, 14, 1082557. <https://doi.org/10.3389/fpsyg.2023.1082557>
- Tal, Z., Geva, R., & Amedi, A. (2016). The origins of metamodality in visual object area LO: Bodily topographical biases and increased functional connectivity to S1. *Neuroimage*, 127, 363-375. <https://doi.org/10.1016/j.neuroimage.2015.11.058>
- Talsma, D. (2015). Predictive coding and multisensory integration: an attentional account of the multisensory mind. *Frontiers in Integrative Neuroscience*, 9, 19. <https://doi.org/10.3389/fnint.2015.00019>

- 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. <https://doi.org/10.1016/j.tics.2010.06.008>
- Tang, X., Fan, D., Wang, X., Xing, Z., Yu, S., Wang, A., & Yu, H. (2025). Exploring how sensory dominance modulated by modality-specific expectation: an event-related potential study. *Frontiers in Psychology*, 16, 1548100. <https://doi.org/10.3389/fpsyg.2025.1548100>
- Tarr, M. J., & Bülthoff, H. H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance*, 21(6), 1494–1505. <https://doi.org/10.1037/0096-1523.21.6.1494>
- Tarr, M. J., & Gauthier, I. (1998). Do viewpoint-dependent mechanisms generalize across members of a class?. *Cognition*, 67(1-2), 73-110. [https://doi.org/10.1016/S0010-0277\(98\)00023-7](https://doi.org/10.1016/S0010-0277(98)00023-7)
- Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21(2), 233-282. [https://doi.org/10.1016/0010-0285\(89\)90009-1](https://doi.org/10.1016/0010-0285(89)90009-1)
- Ten Oever, S., Romei, V., van Atteveldt, N., Soto-Faraco, S., Murray, M. M., & Matusz, P. J. (2016). The COGs (context, object, and goals) in multisensory processing. *Experimental Brain Research*, 234(5), 1307-1323. <https://doi.org/10.1007/s00221-016-4590-z>
- Thelen, A., & Murray, M. M. (2013). The efficacy of single-trial multisensory memories. *Multisensory Research*, 26(5), 483-502. <https://doi.org/10.1163/22134808-00002426>
- Thelen, A., Cappe, C., & Murray, M. M. (2012). Electrical neuroimaging of memory discrimination based on single-trial multisensory learning. *NeuroImage*, 62(3), 1478–1488. <https://doi.org/10.1016/j.neuroimage.2012.05.027>
- Thelen, A., Talsma, D., & Murray, M. M. (2015). Single-trial multisensory memories affect later auditory and visual object discrimination. *Cognition*, 138, 148-160. <https://doi.org/10.1016/j.cognition.2015.02.003>
- Thomas, J. P., & Shiffrar, M. (2013). Meaningful sounds enhance visual sensitivity to human gait regardless of synchrony. *Journal of Vision*, 13(14), 8. <https://doi.org/10.1167/13.14.8>
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582), 520–522. <https://doi.org/10.1038/381520a0>
- Tieri, G., Tidoni, E., Pavone, E. F., & Aglioti, S. M. (2015). Body visual discontinuity affects feeling of ownership and skin conductance responses. *Scientific Reports*, 5(1), 17139. <https://doi.org/10.1038/srep17139>

- Tjan, B. S., & Legge, G. E. (1998). The viewpoint complexity of an object-recognition task. *Vision Research*, 38(15-16), 2335-2350. [https://doi.org/10.1016/S0042-6989\(97\)00255-1](https://doi.org/10.1016/S0042-6989(97)00255-1)
- Tovar, D. A., Murray, M. M., & Wallace, M. T. (2020). Selective enhancement of object representations through multisensory integration. *Journal of Neuroscience*, 40(29), 5604-5615. <https://doi.org/10.1523/JNEUROSCI.2139-19.2020>
- Town, S. M., Wood, K. C., & Bizley, J. K. (2018). Sound identity is represented robustly in auditory cortex during perceptual constancy. *Nature Communications*, 9(1), 4786. <https://doi.org/10.1038/s41467-018-07237-3>
- Tran, T. Q., Langlotz, T., Young, J., Schubert, T. W., & Regensbrecht, H. (2024). Classifying presence scores: Insights and analysis from two decades of the igroup presence questionnaire (ipq). *ACM Transactions on Computer-Human Interaction*, 31(5), 1-26. <https://doi.org/10.1145/3689046>
- Trimble Inc. (2022). *SketchUp Pro* (Version 2022) [Computer software]
- Trommershauser, J., Kording, K., & Landy, M. S. (Eds.). (2011). *Sensory cue integration*. Oxford University Press.
- Tsunada, J., & Cohen, Y. E. (2014). Neural mechanisms of auditory categorization: from across brain areas to within local microcircuits. *Frontiers in Neuroscience*, 8, 161. <https://doi.org/10.3389/fnins.2014.00161>
- Tunik, E., Ortigue, S., Adamovich, S. V., & Grafton, S. T. (2008). Differential recruitment of anterior intraparietal sulcus and superior parietal lobule during visually guided grasping revealed by electrical neuroimaging. *Journal of Neuroscience*, 28(50), 13615-13620. <https://doi.org/10.1523/JNEUROSCI.3303-08.2008>
- Ullman, S. (1979). The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153), 405-426. <https://doi.org/10.1098/rspb.1979.0006>
- Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, 32(3), 193-254. [https://doi.org/10.1016/0010-0277\(89\)90036-X](https://doi.org/10.1016/0010-0277(89)90036-X)
- van Beers, R. J., Sittig, A. C., & Denier van der Gon, J. J. (2002). Integration of proprioceptive and visual position information: An experimentally supported model. *Journal of Neurophysiology*, 81(3), 1355-1364. <https://doi.org/10.1152/jn.1999.81.3.1355>
- Van der Burg, E., Talsma, D., Olivers, C. N., Hickey, C., & Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *Neuroimage*, 55(3), 1208-1218. <https://doi.org/10.1016/j.neuroimage.2010.12.068>

- Van Geert, E., & Wagemans, J. (2023). What good is goodness? The effects of reference points on discrimination and categorization of shapes. *Journal of Experimental Psychology: Human Perception and Performance*, 49(8), 1180–1201. <https://doi.org/10.1037/xhp0001137>
- Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, 15(4), 732–749. <https://doi.org/10.3758/PBR.15.4.732>
- Vanpaemel, W., & Storms, G. (2010). Abstraction and model evaluation in category learning. *Behavior Research Methods*, 42(2), 421–437. <https://doi.org/10.3758/BRM.42.2.421>
- Verrillo, R. T., Fraioli, A. J., & Smith, R. L. (1969). Sensation magnitude of vibrotactile stimuli. *Perception & Psychophysics*, 6(6), 366–372. <https://doi.org/10.3758/BF03212793>
- Viganò, S., Borghesani, V., & Piazza, M. (2021). Symbolic categorization of novel multisensory stimuli in the human brain. *NeuroImage*, 235, 118016. <https://doi.org/10.1016/j.neuroimage.2021.118016>
- Vlach H. A. (2016). How we categorize objects is related to how we remember them: The shape bias as a memory bias. *Journal of Experimental Child Psychology*, 152, 12–30. <https://doi.org/10.1016/j.jecp.2016.06.013>
- Volcic, R., & Camponogara, I. (2018). How do vision and haptics combine in multisensory grasping?. *Journal of Vision*, 18(10), 64–64. <https://doi.org/10.1167/18.10.64>
- Vroomen, J., & de Gelder, B. (2000). Sound enhances visual perception: Cross-modal effects of auditory intensity on visual motion perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26(5), 1545–1556. <https://doi.org/10.1037/0096-1523.26.5.1583>
- Vuilleumier, P., Henson, R. N., Driver, J., & Dolan, R. J. (2002). Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. *Nature Neuroscience*, 5(5), 491–499. <https://doi.org/10.1038/nn839>
- Vuong, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in Cognitive Sciences*, 23(8), 672–685. <https://doi.org/10.1016/j.tics.2019.04.013>
- Vuong, Q. C., & Tarr, M. J. (2004). Rotation direction affects object recognition. *Vision Research*, 44(14), 1717–1730. <https://doi.org/10.1016/j.visres.2004.02.002>
- Vuong, Q. C., & Tarr, M. J. (2006). Structural similarity and spatiotemporal noise effects on learning dynamic novel objects. *Perception*, 35(4), 497–510. <https://doi.org/10.1068/p5491>
- Vuong, Q. C., Friedman, A., & Plante, C. (2009). Modulation of viewpoint effects in object recognition by shape and motion cues. *Perception*, 38(11), 1628–1648. <https://doi.org/10.1068/p6430>

- Vuong, Q. C., Friedman, A., & Read, J. C. (2012). The relative weight of shape and non-rigid motion cues in object perception: A model of the parameters underlying dynamic object discrimination. *Journal of Vision*, 12(3), 16-16. <https://doi.org/10.1167/12.3.16>
- Vuong, Q. C., Peissig, J. J., Harrison, M. C., & Tarr, M. J. (2005). The role of surface pigmentation for recognition revealed by contrast reversal in faces and Greebles. *Vision Research*, 45(10), 1213-1223. <https://doi.org/10.1016/j.visres.2004.11.015>
- Wallace, M. T., Meredith, M. A., & Stein, B. E. (1992). Integration of multiple sensory modalities in cat cortex. *Experimental Brain Research*, 91(3), 484-488. <https://doi.org/10.1007/BF00227844>
- Wallace, M. T., Meredith, M. A., & Stein, B. E. (1998). Multisensory integration in the superior colliculus of the alert cat. *Journal of Neurophysiology*, 80(2), 1006-1010. <https://doi.org/10.1152/jn.1998.80.2.1006>
- Wallraven, C., & Bülthoff, H. H. (2001). Automatic acquisition of exemplar-based representations for recognition from image sequences. In *CVPR 2001 Workshop on Models versus Exemplars in Computer Vision* (pp. 1-9).
- Wallraven, C., & Bülthoff, H. H. (2009). Multimodal Categorization. In *Object Categorization: Computer and Human Vision Perspectives* (pp. 488-501). Cambridge University Press. <https://doi.org/10.1017/CBO9780511635465.027>
- Wallraven, C., Bülthoff, H. H., Waterkamp, S., van Dam, L., & Gaißert, N. (2014). The eyes grasp, the hands see: Metric category knowledge transfers between vision and touch. *Psychonomic Bulletin & Review*, 21(4), 976-985. <https://doi.org/10.3758/s13423-013-0563-4>
- Wallraven, C., Whittingstall, L., & Bülthoff, H. H. (2013). Learning to recognize face shapes through serial exploration. *Experimental Brain Research*, 226(4), 513-523. <https://doi.org/10.1007/s00221-013-3463-y>
- Wang, G., Wang, H. H., & Ren, G. (2024). Visual and Haptic Guidance for Enhancing Target Search Performance in Dual-Task Settings. *Applied Sciences*, 14(11), 4650. <https://doi.org/10.3390/app14114650>
- Wang, Y., & Zeng, Y. (2022). Multisensory concept learning framework based on spiking neural networks. *Frontiers in Systems Neuroscience*, 16, 845177. <https://doi.org/10.3389/fnsys.2022.845177>
- Wang, Z., Yu, L., Xu, J., Stein, B. E., & Rowland, B. A. (2020). Experience creates the multisensory transform in the superior colliculus. *Frontiers in Integrative Neuroscience*, 14, 18. <https://doi.org/10.3389/fnint.2020.00018>

- Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, 88(3), 638–667. <https://doi.org/10.1037/0033-2909.88.3.638>
- Welch, R. B., Dutton-Hurt, L. D., & Warren, D. H. (1986). Contributions of audition and vision to temporal rate perception. *Perception & Psychophysics*, 39(4), 294–300. <https://doi.org/10.3758/BF03204939>
- Wenk, N., Buetler, K. A., Penalver-Andres, J., Müri, R. M., & Marchal-Crespo, L. (2022). Naturalistic visualization of reaching movements using head-mounted displays improves movement quality compared to conventional computer screens and proves high usability. *Journal of NeuroEngineering and Rehabilitation*, 19(1), 137. <https://doi.org/10.1186/s12984-022-01101-8>
- Wertheimer, M. (1938). Laws of organization in perceptual forms. In *A source book of Gestalt Psychology* (pp. 71–88). <https://doi.org/10.1037/11496-005>
- Westfall, J. (2016). PANGAEA: Power ANALysis for GEneral ANOVA designs [Computer software]. Retrieved from <https://jakewestfall.org/pangea/>
- Weydmann, G., Palmieri, I., Simões, R. A., Centurion Cabral, J. C., Eckhardt, J., Tavares, P., ... & Bizarro, L. (2023). Switching to online: Testing the validity of supervised remote testing for online reinforcement learning experiments. *Behavior Research Methods*, 55(7), 3645–3657. <https://doi.org/10.3758/s13428-022-01982-6>
- Wijntjes, M. W., Volcic, R., Pont, S. C., Koenderink, J. J., & Kappers, A. M. (2009). Haptic perception disambiguates visual perception of 3D shape. *Experimental Brain Research*, 193(4), 639–644. <https://doi.org/10.1007/s00221-009-1713-9>
- Williams, J. R., Markov, Y. A., Tiurina, N. A., & Störmer, V. S. (2022). What You See Is What You Hear: Sounds Alter the Contents of Visual Perception. *Psychological Science*, 33(12), 2109–2122. <https://doi.org/10.1177/09567976221121348>
- Witzel, C., & Gegenfurtner, K. R. (2016). Categorical perception for red and brown. *Journal of experimental psychology. Human Perception and Performance*, 42(4), 540–570. <https://doi.org/10.1037/xhp0000154>
- Wong, A. C., Palmeri, T. J., & Gauthier, I. (2009). Conditions for facelike expertise with objects: becoming a Ziggerin expert-but which type?. *Psychological Science*, 20(9), 1108–1117. <https://doi.org/10.1111/j.1467-9280.2009.02430.x>
- Woodcock, J., Davies, W. J., & Cox, T. J. (2017). A cognitive framework for the categorisation of auditory objects in urban soundscapes. *Applied Acoustics*, 121, 56–64. <https://doi.org/10.1016/j.apacoust.2017.01.027>

- Woods, A. T., & Newell, F. N. (2004). Visual, haptic and cross-modal recognition of objects and scenes. *Journal of Physiology-Paris*, 98(1-3), 147-159. <https://doi.org/10.1016/j.jphysparis.2004.03.006>
- Woods, A. T., Levstek, M., Moffatt, J., Lycett, M., Whittaker, L., & Dalton, P. (2025). Going Beyond the Ordinary—User Perceptions of the Impact of Multisensory Elements on Presence in Virtual Reality at the Royal Opera House. *Multisensory Research*, 38(1-3), 1-25. <https://doi.org/10.1163/22134808-bja10141>
- Wozny, D. R., Beierholm, U. R., & Shams, L. (2008). Human trimodal perception follows optimal statistical inference. *Journal of Vision*, 8(3), 24-24. <https://doi.org/10.1167/8.3.24>
- Wu, J., Fu, Q., & Rose, M. (2020). Stimulus modality influences the acquisition and use of the rule-based strategy and the similarity-based strategy in category learning. *Neurobiology of Learning and Memory*, 168, 107152. <https://doi.org/10.1016/j.nlm.2019.107152>
- Wu, J., Li, Q., Fu, Q., Rose, M., & Jing, L. (2021). Multisensory Information Facilitates the Categorization of Untrained Stimuli. *Multisensory Research*, 35(1), 79-107. <https://doi.org/10.1163/22134808-bja10061>
- Yang, Y., Dickey, M. W., Fiez, J., Murphy, B., Mitchell, T., Collinger, J., Tyler-Kabara, E., Boninger, M., & Wang, W. (2017). Sensorimotor experience and verb-category mapping in human sensory, motor and parietal neurons. *Cortex; a journal devoted to the study of the nervous system and behavior*, 92, 304–319. <https://doi.org/10.1016/j.cortex.2017.04.021>
- Yau, J. M., DeAngelis, G. C., & Angelaki, D. E. (2015). Dissecting neural circuits for multisensory integration and crossmodal processing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1677), 20140203. <https://doi.org/10.1098/rstb.2014.0203>
- Yau, J. M., Kim, S. S., Thakur, P. H., & Bensmaia, S. J. (2016). Feeling form: the neural basis of haptic shape perception. *Journal of Neurophysiology*, 115(2), 631-642. <https://doi.org/10.1152/jn.00598.2015>
- Yau, J. M., Olenczak, J. B., Dammann, J. F., & Bensmaia, S. J. (2009). Temporal frequency channels are linked across audition and touch. *Current Biology*, 19(7), 561-566. <https://doi.org/10.1016/j.cub.2009.02.013>
- Yildirim, I., & Jacobs, R. A. (2013). Transfer of object category knowledge across visual and haptic modalities: Experimental and computational studies. *Cognition*, 126(2), 135-148. <https://doi.org/10.1016/j.cognition.2012.08.005>

- Zampini, M., Torresan, D., Spence, C., and Murray, M. M. (2007). Auditory-somatosensory multisensory interactions in front and rear space. *Neuropsychologia* 45, 1869–1877 <https://doi.org/10.1016/j.neuropsychologia.2006.12.004>
- Zangaladze, A., Epstein, C. M., Grafton, S. T., & Sathian, K. (1999). Involvement of visual cortex in tactile discrimination of orientation. *Nature*, 401(6753), 587-590. <https://doi.org/10.1038/44139>
- Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences*, 6(1), 37-46. [https://doi.org/10.1016/s1364-6613\(00\)01816-7](https://doi.org/10.1016/s1364-6613(00)01816-7)
- Zeithamova, D., Maddox, W. T., & Schnyer, D. M. (2008). Dissociable prototype learning systems: evidence from brain imaging and behavior. *Journal of Neuroscience*, 28(49), 13194-13201. <https://doi.org/10.1523/JNEUROSCI.2915-08.2008>
- Zhang, J., Huang, M., Yang, R., Wang, Y., Tang, X., Han, J., & Liang, H. N. (2023). Understanding the effects of hand design on embodiment in virtual reality. *AI EDAM*, 37, e10. <https://doi.org/10.1017/S0890060423000045>
- Zhao, S., Brown, C. A., Holt, L. L., & Dick, F. (2022). Robust and efficient online auditory psychophysics. *Trends in Hearing*, 26,. <https://doi.org/10.1177/23312165221118792>
- Ziat, M. (2023). Haptics for human-computer interaction: From the skin to the brain. *Foundations and Trends® in Human-Computer Interaction*, 17(1-2), 1-194. <http://dx.doi.org/10.1561/11000000061>
- Ziat, M., Jhunjunwala, R., Clepper, G., Kivelson, P. D., & Tan, H. Z. (2022). Walking on paintings: Assessment of passive haptic feedback to enhance the immersive experience. *Frontiers in Virtual Reality*, 3, 997426. <https://doi.org/10.3389/frvir.2022.997426>
- Zou, Z., Zhao, B., Ting, K. H., Wong, C., Hou, X., & Chan, C. C. (2023). Multisensory integration augmenting motor processes among older adults. *Frontiers in Aging Neuroscience*, 15, 1293479. <https://doi.org/10.3389/fnagi.2023.1293479>
- Zwoliński, G., Kamińska, D., Laska-Leśniewicz, A., & Adamek, Ł. (2022). Vibrating tilt platform enhancing immersive experience in vr. *Electronics*, 11(3), 462. <https://doi.org/10.3390/electronics11030462>

8 Appendices

8.1 Appendix A - Ethics Approval letters

Appendices A.1 Ethics Approval letter for online experiments from the School of Psychology, Trinity College Dublin.



Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

F.A.O Fiona Newell

Approval ID: SPREC102020-50

School of Psychology Research Ethics Committee

9th February 2021

Dear Fiona,

The School of Psychology Research Ethics Committee has reviewed your application entitled "The formation of multisensory object categories in children and adults: An Online Investigation", and I am pleased to inform you that it was approved.

Please note that you will be required to submit a completed **Project Annual Report Form** on each anniversary of this approval, until such time as an **End of Project Report Form is submitted** upon completion of the research. Copies of both forms are available for download from the Ethics section of the School website.

Please note that you must be familiar with and adhere to the attached 'Safety Protocol for Adults'.

Adverse events associated with the conduct of this research must be reported immediately to the Chair of the Ethics Committee.

Yours sincerely,

Richard Carson

Chair,

School of Psychology Research Ethics Committee

Appendices A.2 Ethics Approval letter for in person experiments from the School of Psychology, Trinity College Dublin.



Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

F.A.O. Fiona Newell, Martina Seveso, Eimear McKenna, Forat Alahmed, Isabella Devine, Jack Cotter

Approval ID: SPREC092022-05

School of Psychology Research Ethics Committee

2nd November 2022

Dear all,

The School of Psychology Research Ethics Committee has reviewed your application entitled "Multisensory Object Categories Adults In-Person (MOC-AP)", and I am pleased to inform you that it was approved.

Please note that you will be required to submit a completed **Project Annual Report Form** on each anniversary of this approval, until such time as an **End of Project Report Form is submitted** upon completion of the research. Copies of both forms are available for download from the Ethics section of the School website.

Adverse events associated with the conduct of this research must be reported immediately to the Chair of the Ethics Committee.

Yours sincerely,

Richard Carson
Chair,
School of Psychology Research Ethics Committee

Scoil na Siceolaíochta
Dámh na nEolaíochtaí Sóisialta agus Daonna,
Áras an Phiarsaigh,
Coláiste na Tríonóide,
Baile Átha Cliath,
Ollscoil Átha Cliath,
Baile Átha Cliath 2, Éire.

School of Psychology
Faculty of Arts, Humanities and Social Sciences,
Trinity College Dublin,
The University of Dublin,
Dublin 2, Ireland.

+353 1 896 1886
psychology@tcd.ie
www.tcd.ie/psychology

8.2 Appendix B – Experimental Supplementary Materials

Figure S1

The plot displays the odds ratios for discrimination accuracy across object category, derived from the generalized linear mixed model (glmer) in Experiment 1A, for the sound-object category matching task (panel A) and sound-object exemplar matching task (panel B). The vertical black line at odds ratio = 1 serves as the reference. Points represent the estimated odds ratios, while horizontal lines denote the 95% confidence intervals.

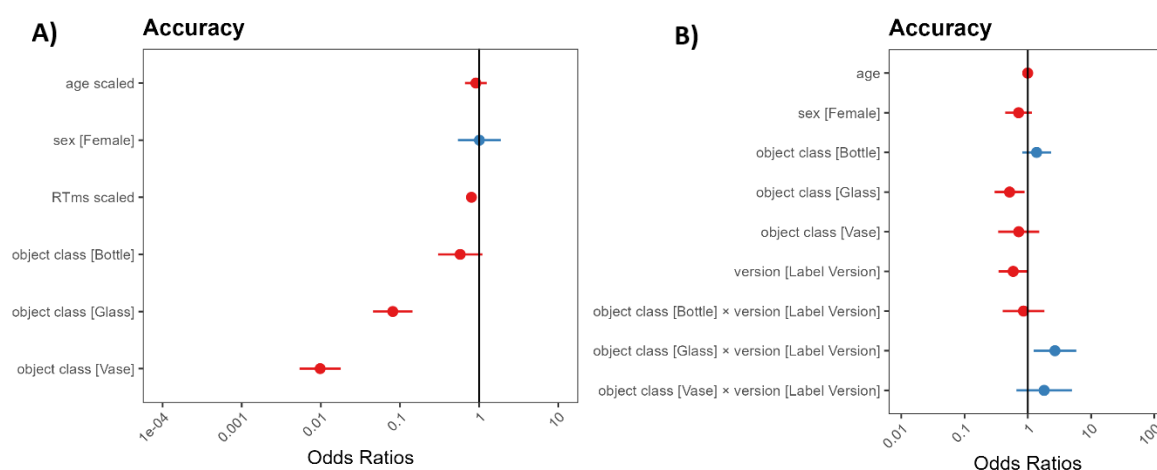


Table S1

Predicted mean accuracy values per object class and version during test (95% confidence intervals) in Experiment 1A.

Sound-object category matching task	
Object Class	
Bell	0.97 (0.95, 0.99)
Bottle	0.96 (0.93, 0.97)
Glass	0.75 (0.68, 0.81)
Vase	0.27 (0.20, 0.34)
Sound-object exemplar matching task	
Version - Image	
Bell	0.55 (0.45, 0.65)
Bottle	0.63 (0.53, 0.72)
Glass	0.39 (0.29, 0.50)
Vase	0.47 (0.31, 0.64)
Version - Label	
Bell	0.42 (0.32, 0.53)
Bottle	0.46 (0.36, 0.56)
Glass	0.50 (0.40, 0.61)
Vase	0.49 (0.35, 0.63)

Table S2

Counterbalancing procedure of objects categories in the 12 groups in Experiment 1B.

P - ID	Group - ID	Learning phase Target objects	Categorical Perception phase Discrimination and Categorisation tasks
1-8	1	Bell (Hand – Church), Bottle (Wine – Coke), Glass (Wine – Beer), Vase (Urn – Single)	<i>Within</i> - Bell, Bottle; <i>Between</i> - wineGlass – urnVase, beerGlass - singlestemVase
9-16	2	Bell (Hand – Church), Bottle (Wine – Coke), Glass (Wine – Beer), Vase (Urn – Stem)	<i>Within</i> - Bell, Bottle; <i>Between</i> - wineGlass – singlestemVase, beerGlass - urnVase
17-24	3	Bell (Hand – Church), Bottle (Wine – Coke), Glass (Wine – Beer), Vase (Urn – Stem)	<i>Within</i> - Bell, Glass; <i>Between</i> – wineBottle – urnVase, cokeBottle - singlestemVase
25-32	4	Bell (Hand – Church), Bottle (Wine – Coke), Glass (Wine – Beer), Vase (Urn – Stem)	<i>Within</i> - Bell, Glass; <i>Between</i> – wineBottle – singlestemVase, cokeBottle - urnVase
33-40	5	Bell (Hand – Church), Bottle (Wine – Coke), Glass (Wine – Beer), Vase (Urn – Stem)	<i>Within</i> - Bell, Vase; <i>Between</i> – wineGlass – cokeBottle, beerGlass - wineBottle
41-48	6	Bell (Hand – Church), Bottle (Wine – Coke), Glass (Wine – Beer), Vase (Urn – Stem)	<i>Within</i> - Bell, Vase; <i>Between</i> – beer Glass – cokeBottle, wineGlass - wineBottle
49-56	7	Bell (Hand – Church), Bottle (Wine – Coke), Glass (Wine – Beer), Vase (Urn – Stem)	<i>Within</i> – Glass, Bottle; <i>Between</i> - handBell – urnVase, churchBell - singlestemVase
57-64	8	Bell (Hand – Church), Bottle (Wine – Coke), Glass (Wine – Beer), Vase (Urn – Stem)	<i>Within</i> - Glass, Bottle; <i>Between</i> - handBell – singlestemVase, churchBell - urnVase
65-72	9	Bell (Hand – Church), Bottle (Wine – Coke), Glass (Wine – Beer), Vase (Urn – Stem)	<i>Within</i> – Glass, Vase; <i>Between</i> – handBell – cokeBottle, churchBell - wineBottle
73-80	10	Bell (Hand – Church), Bottle (Wine – Coke), Glass (Wine – Beer), Vase (Urn – Stem)	<i>Within</i> - Glass, Vase; <i>Between</i> – handBell – wineBottle, churchBell - cokeBottle
81-88	11	Bell (Hand – Church), Bottle (Wine – Coke), Glass (Wine – Beer), Vase (Urn – Stem)	<i>Within</i> - Bottle, Vase; <i>Between</i> – wineGlass – handbell, beerGlass - churchBell
89-96	12	Bell (Hand – Church), Bottle (Wine – Coke), Glass (Wine – Beer), Vase (Urn – Stem)	<i>Within</i> - Bottle, Vase; <i>Between</i> - wineGlass – churchBell, beerGlass - handBell

Table S3

An example of counterbalancing of the learning conditions within participants in Group 3 in Experiment 1B. Object category is defined as within- ('w') and between ('b').

	Learning phase				Categorical Perception phase			
	Exposure			Test	Discrimination		Categorisation	
	AV _c	AV _{inc}	V	AV _{irr}	Visual only			
P 1	Bell (Hand, Church; w)	Vase (Urn, Stem; w)	CokeBot tle - BeerBot tle (b)	‘A’, ‘a’ (b)	Bell (Hand, Church)	Vase (Urn, Stem)	BottleWi ne- GlassWi ne	CokeBot tle - BeerBot tle
P 2	CokeBottl e – BeerBottl e (b)	Bell (Hand, Church; w)	Vase (Urn, Stem; w)	‘B’, ‘b’ (b)	Vase (Urn, Stem)	BottleW ine- GlassWi ne	CokeBot tle - BeerBottl e	Bell (Hand, Church)
P 3	BottleWin e- GlassWin e (b)	CokeBot tle- BeerBot tle (b)	Bell (Hand, Church; w)	‘g’, ‘c’ (w)	Testing set (Learning phase + test objects and sounds)	BottleW ine- GlassWi ne	CokeBot tle - BeerBot tle	Bell (Hand, Church)
P 4	Vase (Urn, Stem; w)	BottleW ine- GlassWi ne (b)	CokeBot tle – BeerBot tle (b)	‘d’, ‘h’ (w)		CokeBo ttle - BeerBot tle	Bell (Hand, Church)	Vase (Urn, Stem)

Table S4

Mean recognition accuracy and reaction time (in milliseconds) for each learning condition and object pairing during the Larning test phase in Experiment 1B.

Learning condition	Target object	Accuracy		Reaction times (ms)	
		Mean	SD	Mean	SD
AVc	beerGlass	1.00	0.00	563.87	125.74
AVc	churchBell	0.95	0.22	723.01	372.99
AVc	cokeBottle	0.97	0.18	715.27	424.58
AVc	handBell	0.93	0.26	607.29	169.06
AVc	singlestemVase	0.93	0.26	698.81	309.89
AVc	urnVase	0.92	0.27	701.12	326.12
AVc	wallClock	0.15	0.36	718.46	501.10
AVc	watchClock	0.90	0.30	720.66	285.23
AVc	wineBottle	0.90	0.31	717.03	311.06
AVc	wineGlass	0.94	0.24	665.80	202.91
AVinc	beerGlass	0.98	0.13	667.38	254.22
AVinc	churchBell	0.94	0.25	636.53	200.52

AVinc	cokeBottle	0.93	0.26	621.96	144.68
AVinc	handBell	0.95	0.22	677.33	230.01
AVinc	highlighterStationary	0.95	0.23	663.63	215.15
AVinc	highlighterStationaty	1.00	0.00	762.75	193.83
AVinc	penStationary	0.97	0.18	669.04	260.92
AVinc	penStationaty	1.00	0.00	656.09	129.89
AVinc	singlestemVase	0.94	0.23	837.13	360.31
AVinc	urnVase	0.94	0.24	726.23	337.46
AVinc	wineBottle	1.00	0.00	690.92	314.44
AVinc	wineGlass	0.97	0.18	647.38	237.11
AVirr	a	0.90	0.31	738.87	225.66
AVirr	A	0.81	0.39	772.57	430.16
AVirr	b	0.94	0.24	721.13	225.99
AVirr	B	0.78	0.42	734.13	270.85
AVirr	c	0.92	0.27	784.02	174.30
AVirr	C	0.87	0.34	749.32	218.67
AVirr	d	0.94	0.24	838.98	312.36
AVirr	D	0.82	0.38	880.62	392.49
AVirr	e	0.67	0.48	909.86	304.82
AVirr	E	0.84	0.37	885.93	969.72
AVirr	g	0.84	0.37	856.62	368.79
AVirr	G	0.94	0.23	741.78	276.41
V	beerGlass	0.97	0.18	661.36	259.35
V	churchBell	0.94	0.24	640.64	172.71
V	cokeBottle	0.98	0.14	627.21	218.91
V	handBell	0.94	0.24	708.68	247.07
V	housePhone	0.97	0.16	679.46	416.54
V	mobilePhone	0.96	0.21	668.95	303.58
V	singlestemVase	0.95	0.23	990.64	1715.08
V	urnVase	0.97	0.18	652.83	147.11
V	wineBottle	0.90	0.31	628.45	297.85
V	wineGlass	0.93	0.25	633.98	166.45

Table S5*Mean categorisation and discrimination accuracy for each object pairing in Experiment 1B.*

Object Pair	Categorisation		Discrimination	
	Mean	SD	Mean	SD
beerGlass_churchBell	0.49	0.50	0.91	0.29
beerGlass_cokeBottle	0.44	0.50	0.87	0.34
beerGlass_handBell	0.42	0.50	0.80	0.40
beerGlass_wineBottle	0.32	0.47	0.92	0.27
Bell	0.45	0.50	0.84	0.36
Bottle	0.42	0.49	0.68	0.47
churchBell_cokeBottle	0.44	0.50	0.82	0.38
churchBell_singlestemVase	0.46	0.50	0.76	0.43
churchBell_wineBottle	0.44	0.50	0.88	0.33
cokeBottle_singlestemVase	0.53	0.50	0.76	0.43
wineGlass_cokeBottle	0.37	0.48	0.85	0.36
Glass	0.45	0.50	0.69	0.46
handBell_cokeBottle	0.47	0.50	0.76	0.43
handBell_wineBottle	0.49	0.50	0.71	0.45
wineGlass_handBell	0.47	0.50	0.91	0.29
beerGlass_singlestemVase	0.47	0.50	0.81	0.39
singlestemVase_handBell	0.55	0.50	0.64	0.48
singlestemVase_wineBottle	0.40	0.49	0.81	0.39
beerGlass_urnVase	0.19	0.40	0.78	0.41
urnVase_churchBell	0.36	0.48	0.82	0.38
urnVase_cokeBottle	0.40	0.49	0.72	0.45
urnVase_handBell	0.49	0.50	0.84	0.36
wineBottle_urnVase	0.22	0.42	0.87	0.33
Vase	0.43	0.50	0.84	0.36
wineGlass_wineBottle	0.34	0.48	0.85	0.36
wineGlass_churchBell	0.51	0.50	0.78	0.41
wineGlass_singlestemVase	0.57	0.50	0.78	0.41
wineGlass_urnVase	0.60	0.49	0.78	0.42

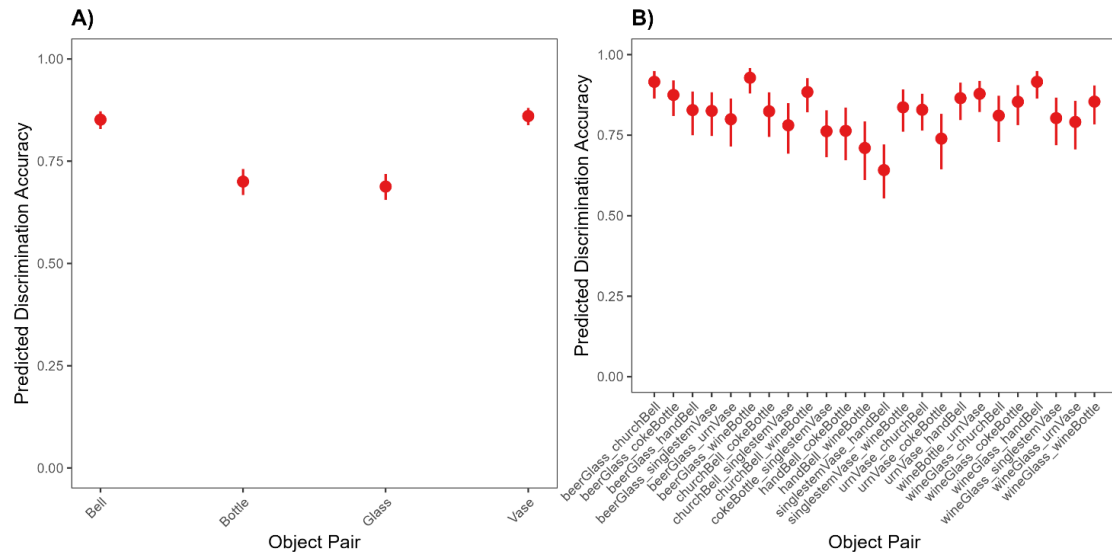
Table S6

Mean responses times (milliseconds) during categorisation and discrimination tasks for each object pairing in Experiment 1B.

Object pairing	Categorisation		Discrimination	
	Mean	SD	Mean	SD
beerGlass_churchBell	765	394	808	275
beerGlass_cokeBottle	804	384	793	279
beerGlass_handBell	684	356	876	391
beerGlass_wineBottle	708	383	692	219
Bell	671	336	750	298
Bottle	701	359	873	386
churchBell_cokeBottle	722	257	786	266
churchBell_singlestemVase	733	376	719	355
churchBell_wineBottle	661	320	894	325
cokeBottle_singlestemVase	696	272	877	283
wineGlass_cokeBottle	688	378	755	280
Glass	707	346	943	414
handBell_cokeBottle	642	290	935	369
handBell_wineBottle	829	382	956	375
wineGlass_handBell	682	337	731	222
beerGlass_singlestemVase	738	295	856	393
singlestemVase_handBell	707	293	1055	491
singlestemVase_wineBottle	715	284	873	434
beerGlass_urnVase	603	291	837	326
urnVase_churchBell	632	275	869	330
urnVase_cokeBottle	807	368	866	412
urnVase_handBell	845	575	763	413
wineBottle_urnVase	627	193	794	223
Vase	712	374	774	290
wineGlass_wineBottle	823	488	839	303
wineGlass_churchBell	674	332	786	390
wineGlass_singlestemVase	650	339	770	331
wineGlass_urnVase	663	256	823	280

Figure S2

The plot shows discrimination accuracy estimation for the within- (panel A), and between-category object pairings (panel B) predicted from the linear mixed-effects model for object pairings in Experiment 1B. Error bars represent 95% confidence intervals.

**Figure S3**

Fixed effect estimates for discrimination accuracy in Experiment 1B for categorical perception analysis. This plot presents fixed effect ($\pm 95\%$ confidence intervals) from the final linear mixed-effects model predicting discrimination accuracy based on Object pairings. Panel (A) refers to visual-only learning condition, panel (B) refers to audio-visual learning conditions.

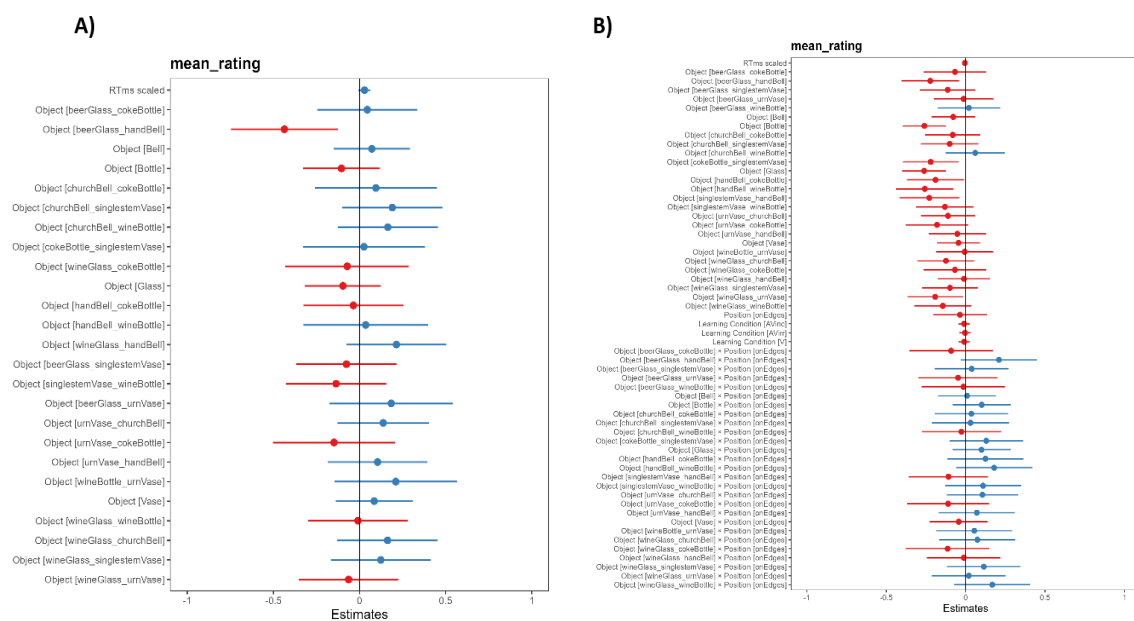
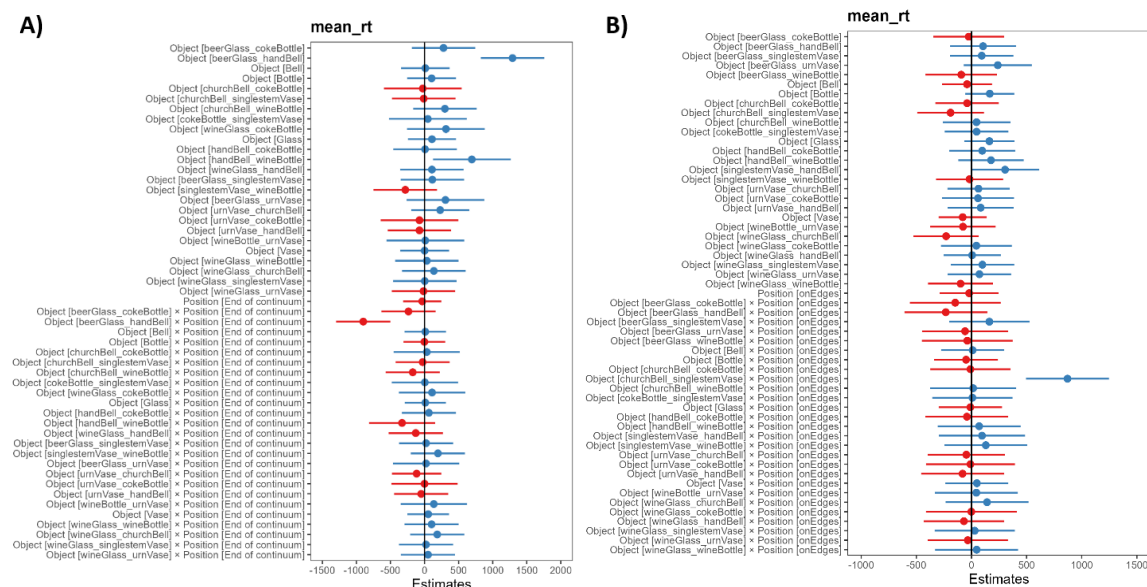


Figure S4

Fixed effect estimates for discrimination reaction times in Experiment 1B for categorical perception analysis. This plot presents fixed effect ($\pm 95\%$ confidence intervals) from the final linear mixed-effects model predicting discrimination reaction times based on Object pairings and position. Panel (A) refers to visual-only learning condition, panel (B) refers to audio-visual learning conditions.

**Figure S5**

The plot shows the predicted reaction times from the linear mixed-effects model across learning conditions and object pairings, during categorisation (panel A) and discrimination (panel B) tasks in Experiment 1B. Learning conditions: audio-visual semantic congruent (AVC) in green; audio-visual semantic incongruent (AVINC) in red; audio-visual semantic irrelevant (AVIRR) in yellow; and unimodal visual only (V) in blue. Error bars represent 95% confidence intervals.

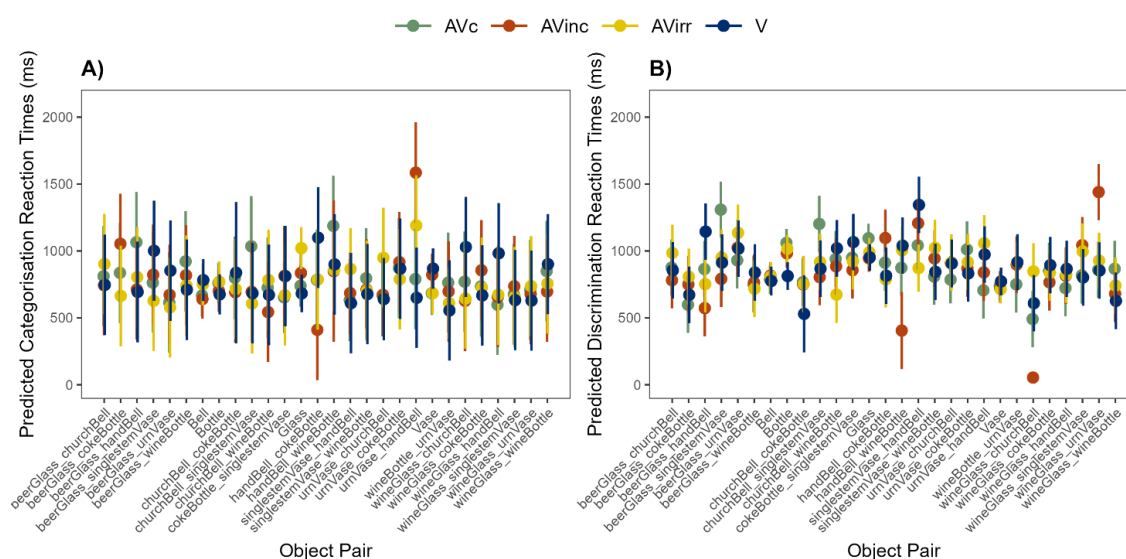
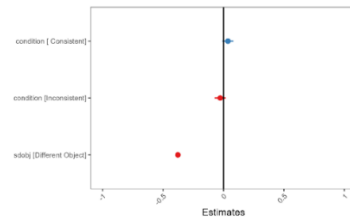


Figure S5

Fixed effect estimates for similarity ratings in Experiment 2A. This plot presents fixed effects ($\pm 95\%$ confidence intervals) from the final linear mixed-effects model predicting similarity ratings based on motion condition (reference = Static) and object similarity (reference = Same Object).



S6. NMDS stress plots and clustering by condition in Experiment 2A. Non-metric multidimensional scaling (NMDS) was applied to similarity matrices from the Static, Consistent, and Inconsistent motion conditions ($1 - r$, Pearson correlation) using the *smacof* package with ordinal constraints. Scree plots of stress values (1–6 dimensions) showed a clear elbow at two dimensions, which was retained for all conditions (Figure S6, top row). Stress values were low and stable across distance transformations (primary: Static = 0.012, Consistent = 0.008, Inconsistent = 0.012; Lawler: 0.048, 0.040, 0.050). K-means clustering of the two-dimensional solutions, evaluated with the Calinski–Harabasz index (NbClust package), indicated four clusters in the Static and Inconsistent conditions and three in the Consistent condition (Figure S6, bottom row).

Figure S6 Top row: scree plots (stress vs. dimensions) for Static (a), Consistent (b), and Inconsistent (c) conditions. Middle row: nMDS configurations across conditions. Bottom row: cluster solutions (black points) with centroids (red crosses), illustrating how perceptual similarity structure varies across motion conditions.

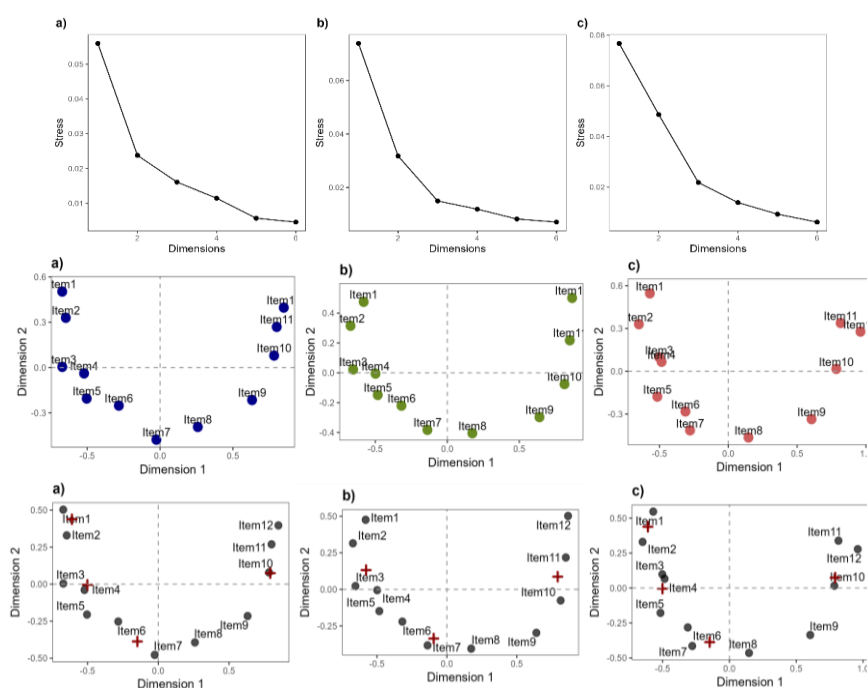
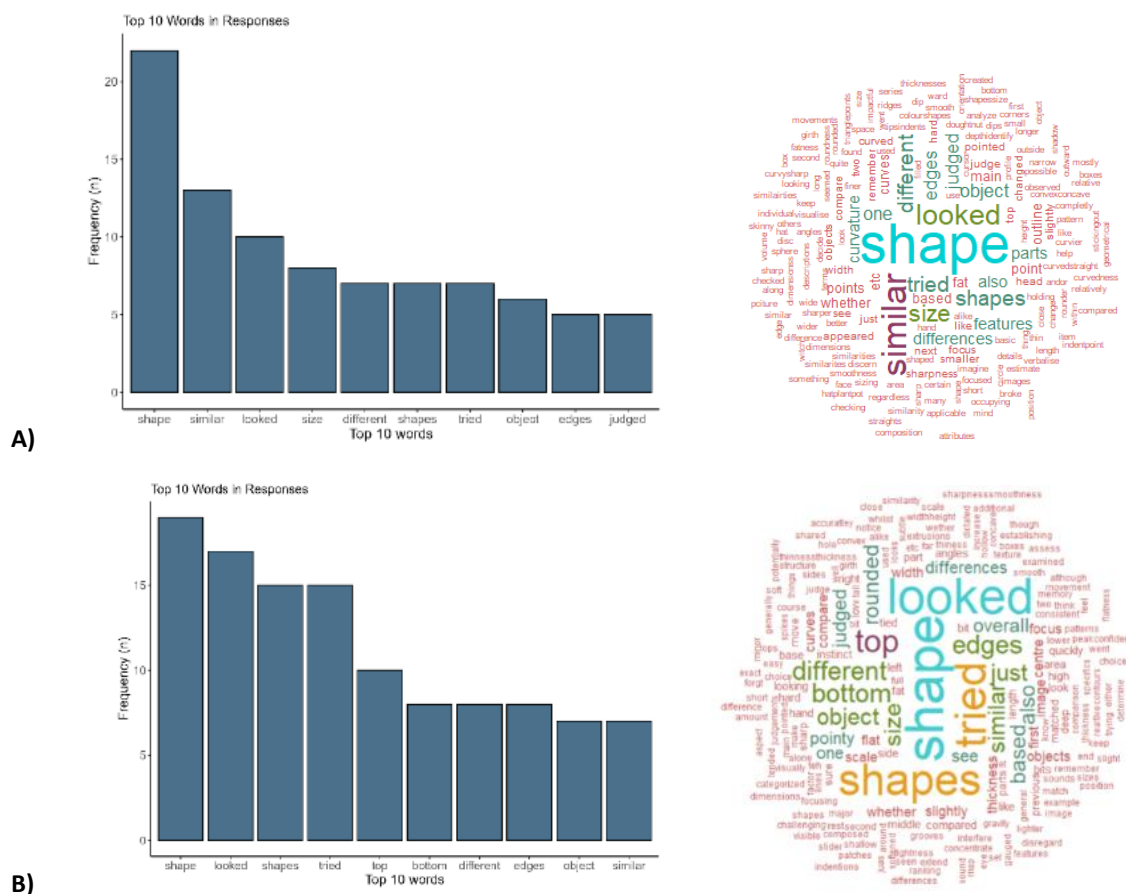


Figure S7

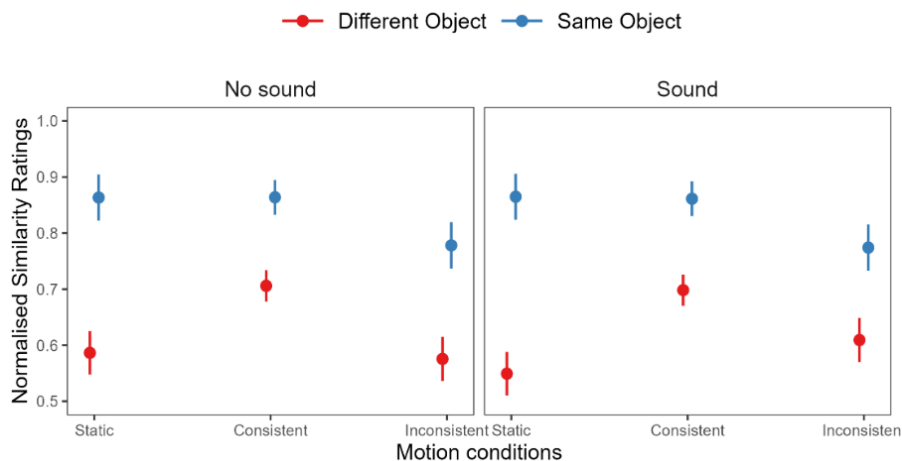
The plot shows the top 10 words used when participants reported their shape similarity judgment strategy alongside a full word cloud in Experiment 2A (panel A) and 2B (panel B).



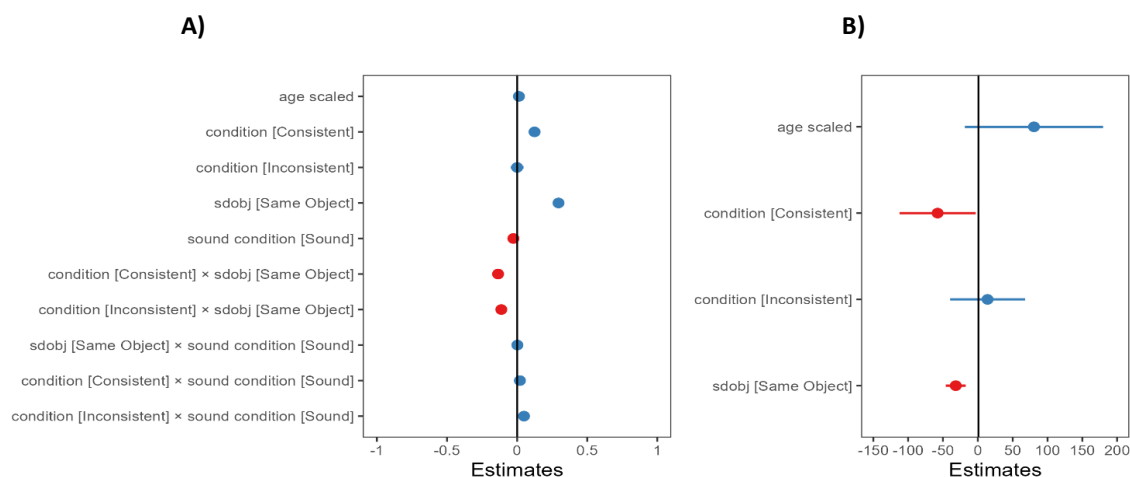
S8 Sound validation study for Experiment 2B. The auditory stimuli were validated through an online two-alternative forced-choice (2AFC) matching task. Seven participants first completed a brief practice session to familiarise themselves with the task structure, followed by the main validation phase. In each trial of the main task, participants listened to a single sound and were asked to identify the corresponding moving pattern. They were instructed to pay close attention, as each sound was presented only once. Each trial began with a fixation cross displayed for 500 ms, followed by the auditory stimulus for 2 seconds. Subsequently, two motion patterns were presented simultaneously for 2 seconds. Participants were then asked to select the object they believed corresponded to the sound, by using the keyboard and pressing the 'A' key to select the object displayed on the left and the 'L' key to select the object on the right. They were instructed to respond as accurately as possible. All auditory stimuli were identified well above chance, with mean accuracies ranging from 0.75 to 1.00. These results confirm that participants were able to reliably associate the sounds with their corresponding motion patterns.

Figure S9

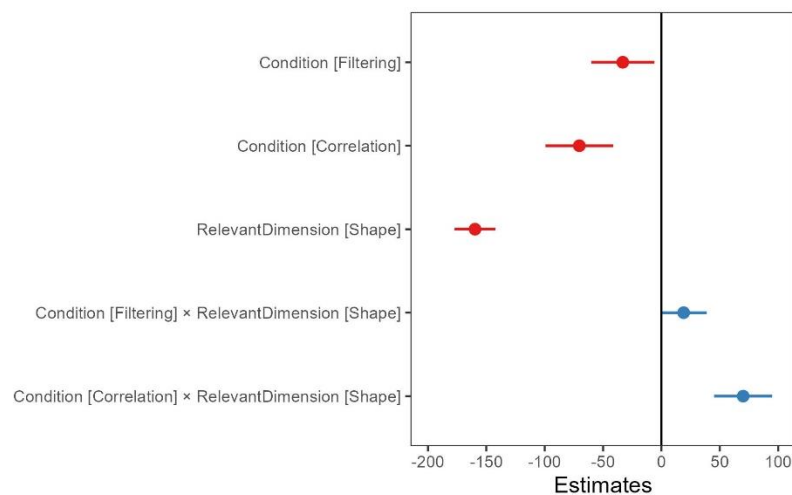
The plot shows normalised similarity ratings predicted from the linear mixed-effects model in Experiment 2B. The model shows a 3-way interaction effect of sound conditions * object pairing * motion condition: in same object pairings (blue) consistently reports higher similarity ratings than different object pairings (red) across all motion and sound conditions. Error bars representing standard errors.

**Figure S10**

Fixed effect estimates ($\pm 95\%$ confidence intervals) from a linear mixed-effects model predicting similarity ratings (panel A) and reaction times (panel B) in Experiment 2B: references are for motion condition (reference = Static), object pair (reference = Different Object), sound condition (reference = No Sound). Positive values indicate increased similarity ratings relative to the reference level.

**Figure S11**

Fixed effect estimates ($\pm 95\%$ confidence intervals) from a linear mixed-effects model predicting reaction times in Experiment 3, based on condition (reference = Baseline), relevant dimension (reference = Motion), and their interactions. Positive values indicate increased reaction times relative to the reference level.



S12 *Exploratory analyses of generalisation performance for Experiment 4 (A and B) and 5 (A and B).*

When interpreting multisensory benefits in general, it is interesting to understand if benefits simply reflect the additive effects of unisensory signals, or whether effects reflect a true *enhancement* afforded by multisensory integration. Importantly, a distinction is observed between the additive effects of unisensory channels and multisensory integration (Murray et al., 2019; Otto et al., 2012) with the latter often leading to better performance than is predicted by probability summation alone, which assumes that performance is influenced by the independent contribution of each sensory cue (see e.g. Stein & Meredith, 1993; Stanford & Stein, 2007). We were therefore interested to explore whether the multisensory benefits observed in the current studies were predicted by probability summation alone or whether there was evidence for cue integration. To this end, we explored whether accuracy improvements under multisensory conditions reflected cue integration or probability summation of independent cues. Previous research has shown that Signal Detection Theory (SDT) provides a good framework for modelling summation (Green & Swets, 1966). Under SDT, component mechanisms give continuous responses that are subject to internal noise, and detection occurs when the maximum response across channels exceeds a criterion level (Baldwin & Schmidtman, 2015). We employed the associated Probability Summation Model (PSM) as the theoretical framework to estimate the expected performance under the assumption that each sensory cue (Shape, Visual Motion, Tactile Motion) operates independently (Meese & Summers, 2007; Baldwin & Schmidtman, 2015; Prins, 2016). We applied this approach to the generalisation data to the unlearned exemplars only (from Experiments 4A, 4B, 5A and 5B) to ensure that performance was not influenced by learning or familiarity effects.

Data Preprocessing: Data were pre-processed to exclude data sets from any participants which were incomplete (e.g., missing Hits or False Alarms), ensuring robust signal detection estimates resulting in a sample size of $N = 30$ for this analysis for each Experiment (i.e., 4A, 4B, 5A, 5B). To control for sampling bias and ensure comparability of multisensory integration across tasks, we also applied a balancing procedure. First, participants data were randomly shuffled to eliminate order effects. Then, equal numbers of participants were sampled from each version of the experiment (blocked or interleaved) for the cue conditions of interest. These subsets were then merged with the remaining cue conditions, $S_v M_v$, $S_v M_t$ and $S_v M_{vt}$, creating a complete and representative dataset for analysis.

Signal Detection Analysis: Accuracy scores were transformed into Hit Rates (H) and False Alarm Rates (F) as follows: Hit Rate (H) was defined as the proportion of correct categorisations for the target category, we considered target category the Category A (e.g., a category A stimulus is correctly identified as belonging to category A – a category A hit); False Alarm Rate (F) was defined as proportion of incorrect categorisations for the non-target category (e.g., a category B stimulus is incorrectly identified as belonging to category A – a category A false alarm). These proportions were bound within [0.01 - 0.99] to ensure stable z-transformation, following standard Signal Detection Theory (SDT) practices (Green & Swets, 1966). The discriminability index (d') was calculated using the SDT formula: $d' = z(H) - z(F)$ where z represents the inverse cumulative distribution function of the normal distribution. d' was computed for each cue condition: S_v , $S_v M_v$, $S_v M_t$, $S_v M_{vt}$.

Model Predictions: Two theoretical models were used to generate expected d' values under different assumptions:

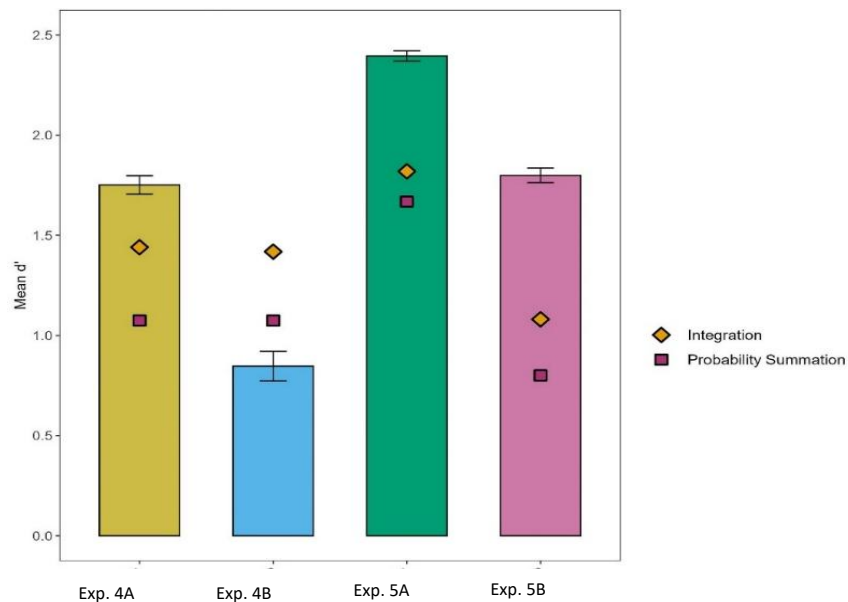
1. *Probability Summation:* Based on the assumption of independent detection across sensory channels, the predicted d' values were computed using a conservative approximation, taking the maximum of the component unisensory d' values (Green & Swets, 1974): $d'_{\text{summation}} = \max(d'_1, d'_2)$. This model reflects the assumption that observers respond correctly when either of the available cues alone suffices for a correct judgement (Garcia-Perez, 1998; Meese & Summers, 2007; Schmidtman et al., 2015).
2. *Integration Model:* Based on optimal cue integration principles, the predicted d' values were calculated using the root-sum-of-squares (Pythagorean summation) of unisensory d' scores: $d'_{\text{integrated}} = \sqrt{(d'_1)^2 + (d'_2)^2}$. This approach is commonly used in multisensory research to model statistically

optimal integration of independent sensory cues (e.g., Alais & Burr, 2004; Ernst & Banks, 2002). It assumes that each cue provides independent and additive information, and that the brain combines them to maximise perceptual precision.

We then compared the observed d' in the S_vM_{vt} (i.e., three-cue) condition to the generalisation predictions made from each of these models. In summary, across both versions of Experiments 4 and 5, the observed d' values significantly differed from the predictions of both the integration and probability summation models but were not consistently greater than predicted (see Figure S12). For Experiment 4A (blocked trials), the observed performance significantly exceeded both models, suggesting evidence for super-additive integration ($z = 6.75, p < .001$; $z = 14.70, p < .001$, respectively for integration and probability summation models). In contrast, in Experiment 4B (interleaved trials), the observed performance was significantly lower than the predictions from either model ($z = -7.78, p < .001$; $z = -3.10, p = .002$, respectively for integration and probability summation models), suggesting that interleaved cue presentation disrupted effective cue integration. In Experiment 5 (reduced shape cue), the observed d' values consistently and substantially exceeded the predictions from both models across both blocked (Experiment 5A) and interleaved (Experiment 5B) conditions. In Experiment 5A, the observed d' was significantly higher than that predicted by the integration model ($z = 22.59, p < .001$) and the probability summation model ($z = 28.51, p < .001$). Similarly, in Experiment 5B, the observed performance was greater than that predicted by both models ($z = 19.92, p < .001$; $z = 27.68, p < .001$, respectively for integration and probability summation models). Visualisations of the model fits revealed that the observed performance often exceeded the performance predicted by probability summation in the S_vM_{vt} condition, indicating that the combination of cues led to performance enhancements consistent with multisensory integration.

Figure S12

Generalisation performance (d') for the S_vM_{vt} condition for the unlearned object exemplars in both the blocked and interleaved versions of Experiment 4 and 5, and the predictions from each of the models.



Note: Each bar represents the observed d' in each experiment, with error bars showing ± 1 standard error of the mean. The overlaid symbols represent the predicted d' values from the Integration model (diamonds) and the Probability Summation model (squares). See text for details.

Discussion: In the blocked trial versions of Experiment 4 and 5, and the interleaved version of Experiment 5, participants' generalisation performance significantly exceeded the predictions of both models, suggesting super-additive integration, i.e. performance gains beyond what would be expected from probabilistic cue summation. This finding supports models of cue integration (e.g., Ernst & Banks, 2002; Ma et al., 2006), in which reliable cues to categorisation are combined to enhance performance. Notably the observed performance exceeded also the optimal integration value, which could indicate either strategic amplification of redundant information or attentional enhancement when cue structure is stable and predictable. In contrast to these findings is the performance to the interleaved trials in Experiment 4B. In this version of the experiment, participants' performance was significantly lower than predicted by both models. This result suggesting that the uncertain nature of the availability of cues from trial to trial affects the integration process. Interleaved presentation might impose higher cognitive demands, preventing participants from reliably selecting which cues to attend for integration (Rohe & Noppeney, 2015), resulting in cue switching across trials (Whiteley & Sahani, 2008). However, in a seemingly contradictory manner,

performance to interleaved trials in Experiment 5 was greater than that predicted by the models. In this experiment, shape similarity was reduced as a reliability cue for categorisation. Consequently, we believe this suggests that participants adapted by reweighting cues, relying more heavily on motion (both visual and tactile) than shape, in a manner consistent with reliability weighting models of integration (e.g. Alais & Burr, 2004; Körding et al., 2007).

Table S7

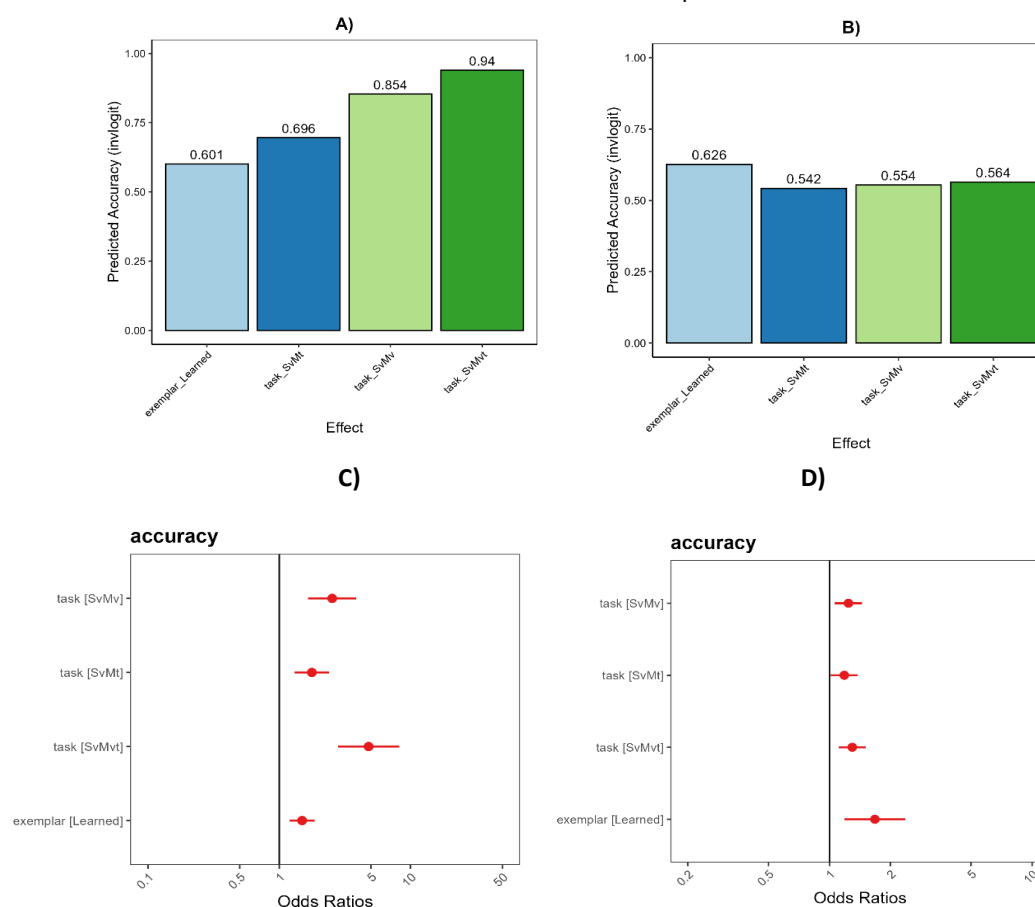
Synchrony assessment task analysis and results across Experiments 4 (A and B), 5 (A and B), and 6.

To assess perceived synchrony across three stimulus onset asynchronies (SOAs: 0 ms, 150 ms, 300 ms), we conducted separate Welch ANOVAs for each experiment. Assumption checks revealed violations of normality and homogeneity of variance in all experiments except Experiment 6 (see Table below). Accordingly, Welch ANOVA was used due to its robustness to such violations.

Experiment	Shapiro–Wilk (Normality)	Levene’s Test (Homogeneity)	Welch ANOVA (<i>F</i> , <i>df</i>)	<i>p</i>	Significant Pairwise Differences
<i>Exp. 4A</i>	$W = .910, p < .001$	$F(2, 483) = 3.85, p = .022$	6.12 (2, 86.87)	.003	0ms–150 ms ($p = .016$) 0ms–300 ms ($p = .005$) 150ms–300 ($p = .70$, ns)
<i>Exp. 4B</i>	$W = .937, p < .001$	$F(2, 346) = 4.72, p = .009$	13.54 (2, 61.83)	<.001	0ms–150 ms ($p < .0001$) 0ms–300 ms ($p = .0003$) 150ms–300ms ($p = 0.84$, ns)
<i>Exp. 5A</i>	$W = .930, p < .001$	$F(2, 343) = 8.17, p < .001$	5.80 (2, 63.88)	.005	0ms–300 ms ($p = .003$) 0ms–150 ms($p=.078$, ns) 150ms–300ms($p=.46$,ns)
<i>Exp. 5B</i>	$W = .923, p < .001$	$F(2, 344) = 6.39, p = .002$	4.82 (2, 63.97)	.011	0ms–150 ms ($p = .024$) 0ms–300 ms ($p = .025$) 150ms–300ms($p=.99$,ns)
<i>Exp. 6</i>	$W = .933, p < .001$	$F(2, 345) = 2.53, p = .080$ (ns)	5.62 (2, 61.76)	.006	0ms–150 ms ($p = .025$) 0ms–300 ms ($p = .006$) 150ms–300ms($p=.80$,ns)

Figure S13.

Predicted categorisation accuracy by cue condition and exemplar type for Experiment 4A (Panel A) and Experiment 4B (Panel B), based on a binomial mixed-effects model. Accuracy is shown separately for learned and novel exemplars across four cue conditions (Sv, SvMv, SvMt and SvMvt). Points represent model-estimated accuracy with 95% confidence intervals. Odds ratios are shown for Experiment 4A (Panel C) and Experiment 4B (Panel D). Points indicate estimated odds ratios; horizontal lines show 95% confidence intervals. The vertical reference line at OR = 1 reflects chance-level performance.

**Figure S14**

Predicted categorisation accuracy by cue condition and exemplar type for Experiment 5A (Panel A, blocked presentation) and Experiment 5B (Panel B, interleaved presentation), based on a binomial mixed-effects model. Odds ratios are shown for Experiment 5A (Panel C) and Experiment 5B (Panel D). Estimated odds ratios for cue conditions by exemplar types (Panel D) in Experiment 5A. Each point indicates the estimated odds ratio for a given contrast between cue conditions and exemplar types (Sv, SvMv, SvMt, SvMvt). Horizontal lines show 95% confidence intervals; the dashed line at OR = 1 marks the baseline (no difference).

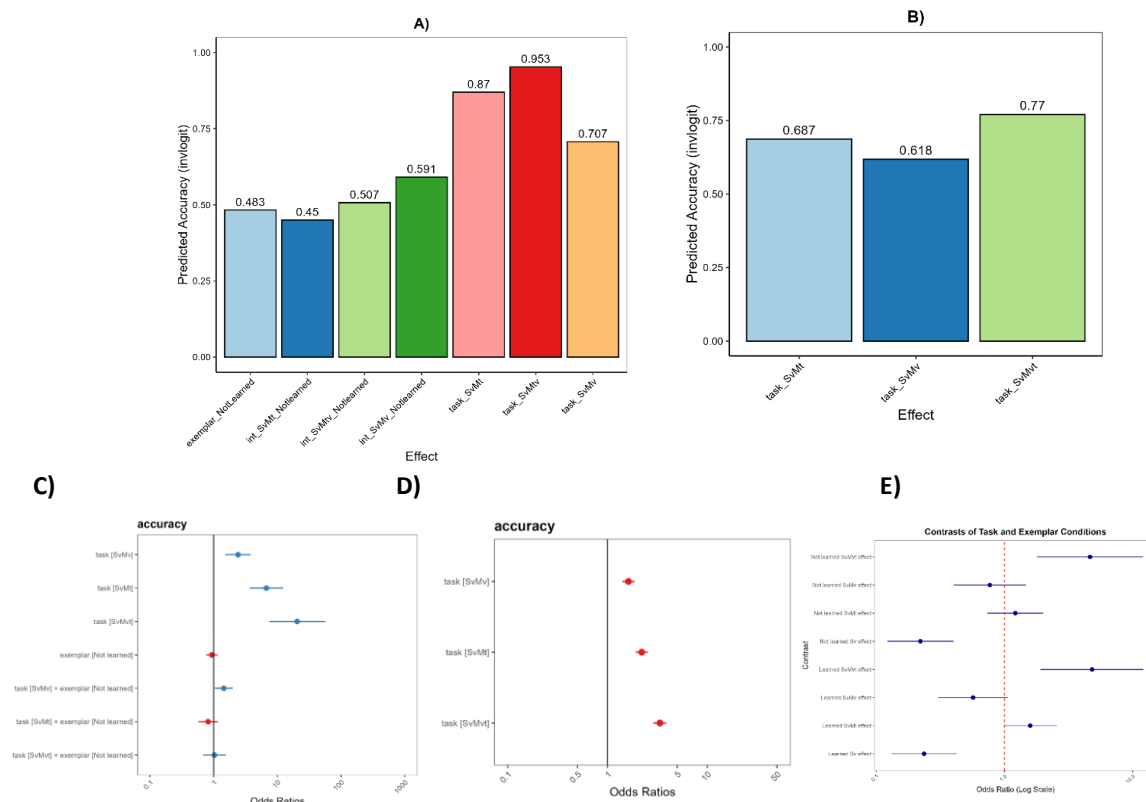
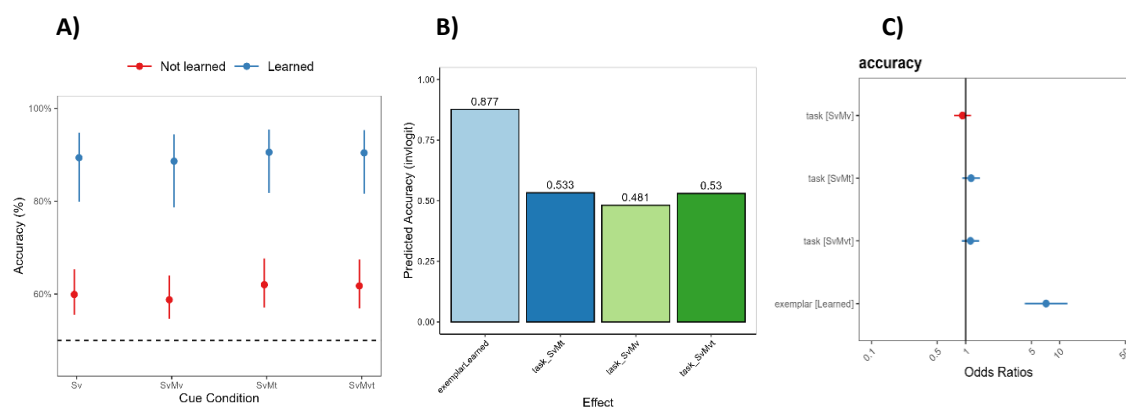


Figure S15

The plot shows the predicted categorisation accuracy estimates (panel A), on the logit scale (panel B) and odds ratios (panel C) for categorisation accuracy by Cue Condition (S_v , S_vM_v , S_vM_t and S_vM_{vt}) and Exemplar Type (learned vs. novel) for Experiment 6. Odds ratios (Panel C) with 95% confidence intervals values above 1 reflect improved accuracy relative to baseline (OR = 1) indicate no effect.



S16 Reaction times analysis for Experiment in Chapter 4. *Experiment 4A:* Likelihood ratio tests predicting RTs indicated a statistically significant main effect of cue condition ($\chi^2(3) = 270$, $p < .001$) but not exemplar type ($\chi^2(1) = 2.66$, $p = .10$), or cue condition by exemplar type interaction ($\chi^2(3) = 0.12$, $p = .99$).in

the model (see FigureS15.A). Post hoc contrast indicated that RTs were significantly faster in the S_v condition compared to S_vM_v (-251 ms, 95% CI [-361, -140], $p < .0001$), S_vM_t (-751 ms, 95% CI [-861, -641], $p < .0001$), and S_vM_{vt} (-654 ms, 95% CI [-755, -535], $p < .0001$). Moreover, RTs were faster in the S_vM_v condition than in both S_vM_t (-501 ms, 95% CI [-611, -390], $p < .0001$) and S_vM_{vt} (-394 ms, 95% CI [-505, -284], $p < .0001$). No significant difference was observed between S_vM_t and S_vM_{vt} (106 ms, 95% CI [-3.96, 217], $p = .066$).

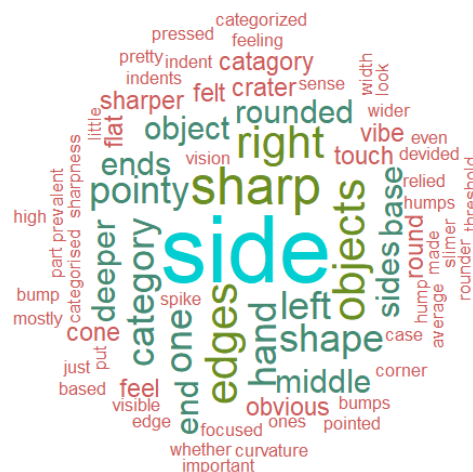
Experiment 4B: Likelihood ratio tests predicting RTs indicated absence of statistically significant main effect of cue condition ($\chi^2(3) = 6.57$, $p = .09$), exemplar type ($\chi^2(1) = 0.26$, $p = .61$) and their interaction was also not statistically significant ($\chi^2(3) = 0.99$, $p = .80$). *Experiment 6:* Likelihood ratio tests revealed significant main effects of cue condition ($\chi^2(3) = 91.4$, $p < .001$) and exemplar type ($\chi^2(1) = 5.29$, $p = .02$), but no significant cue condition by exemplar type interaction ($\chi^2(3) = 0.81$, $p = .85$), see FigureS15.B. Post hoc analyses indicated that RTs were faster in the S_v condition compared to S_vM_v (-120.6 ms, 95% CI [-361, -140], $p = .0004$), S_vM_t (-244 ms, 95% CI [-861, -641], $p < .0001$), and S_vM_{vt} (-278 ms, 95% CI [-755, -535], $p < .0001$). RTs in the S_vM_v condition were faster than S_vM_t (-123 ms, 95% CI [-611, -390], $p = .0002$) and S_vM_{vt} (-158 ms, 95% CI [-505, -284], $p < .0001$). No difference was found between S_vM_t and S_vM_{vt} (-34 ms, 95% CI [-3.96, 217], $p = 1$). Importantly, RTs were faster for learned object exemplars compared to novel exemplars (-47.8 ms, 95% CI [10.4, 14.7], $p = .02$), indicating an effect of prior learning on object categorisation.

Table S8

Questionnaire provided post experiment to assess participants subjective experience in the VR environment and their categorisation strategy in Experiment 7.

Name of the Questionnaires	Description
Simulator Sickness Questionnaire, SSQ Kennedy et al. (1993)	Measures nausea, oculomotor strain, and disorientation
Virtual Embodiment Questionnaire, VEQ Roth & Latoschik (2020)	Measures self-reported ownership, agency, perceived bodily changes.
Igroup Presence Questionnaire, IPQ	Assesses spatial presence (SP), involvement (INV), realism (REAL), and general presence (GP), with reverse-scored items adjusted prior to the calculation of total scores using the IPQ Cal tool.

The plot shows the top 10 words used when participants reported their categorisation strategy alongside a full word cloud in Experiment 7.



The plots show the category representation in Block1 (panel A) and Block 2 (panel B) for each participant. Block 2 refers to no conflict trials only in Experiment 7.

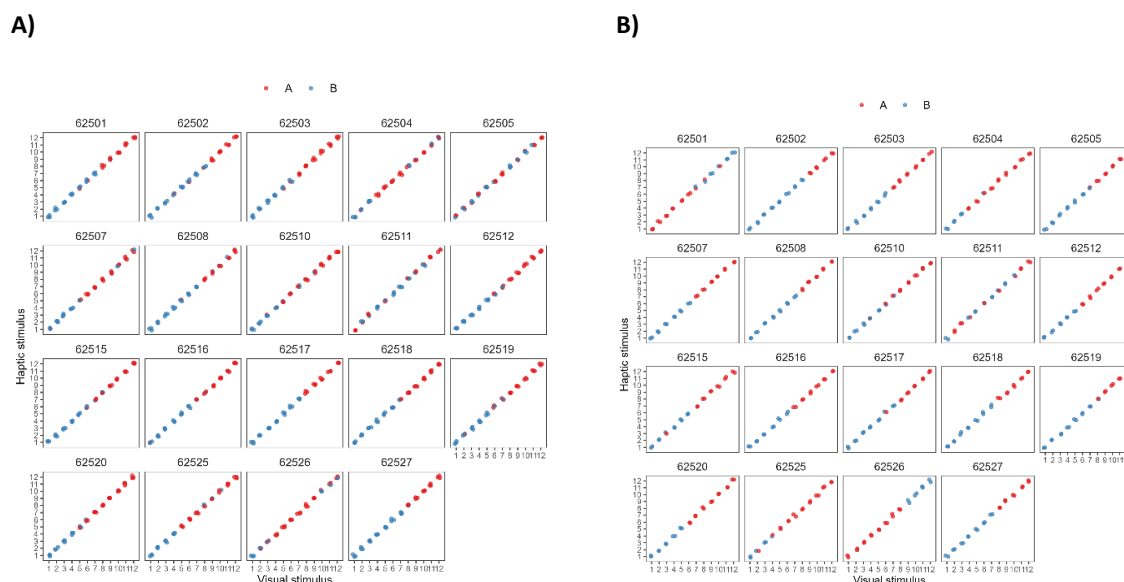
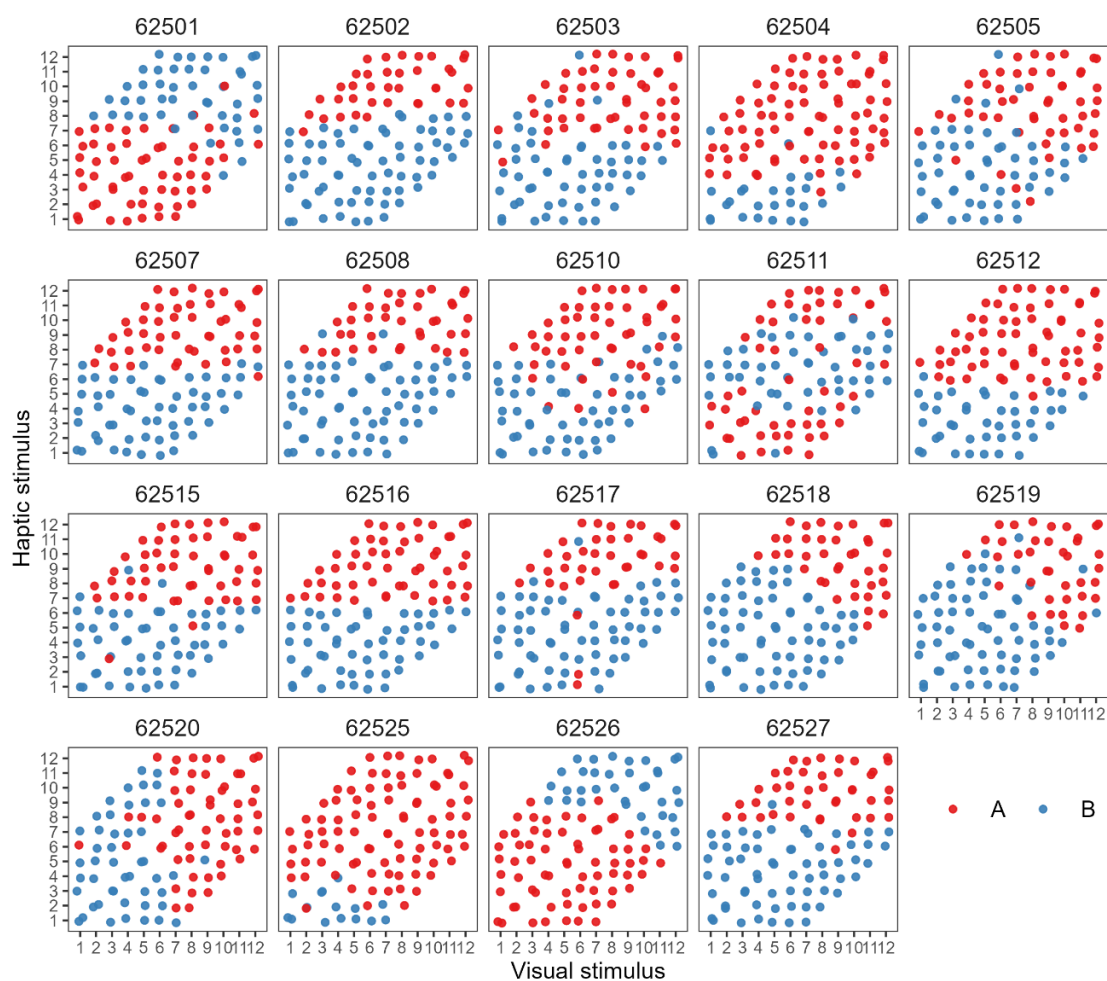


Figure S19

The plot shows the category representation in Block 2 (across all conflict levels) for each participant. Block 2 in Experiment 7.

**Figure S20**

The plot shows the proportion of categorisation responses matching the haptic category (red) or visual category (blue) as a function of conflict level (2–6) for each stimulus set in Experiment 7. Solid horizontal lines represent chance level (.50), and dashed horizontal lines represent the expected proportion under random responding (.25).

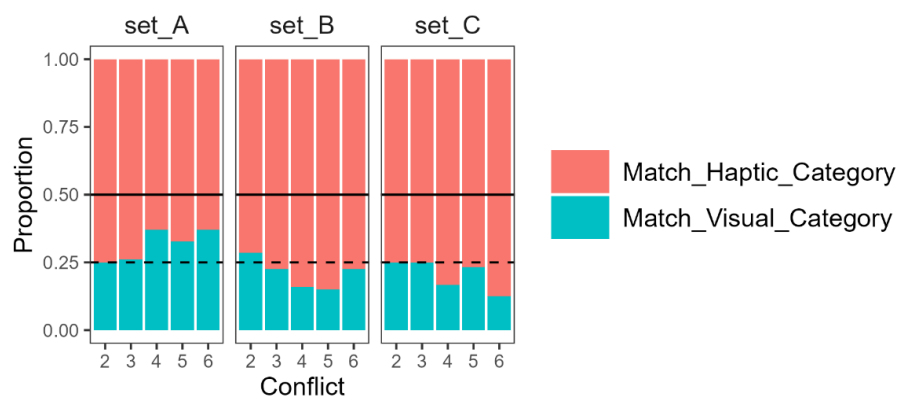


Figure S21

The plot shows the proportions of responses across conflict levels for each participant, and the relative classification based on their reliance on either the haptic (red) or visual (blue) shape cue in Experiment 7.

