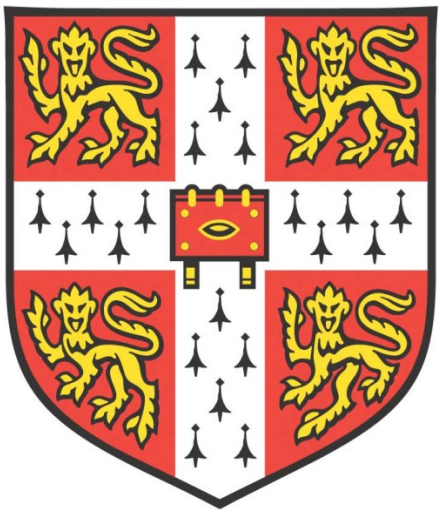
Concepts and Schemas: Representational Format for Structured Knowledge



Levan Bokeria

Hughes Hall

MRC Cognition and Brain Sciences Unit

University of Cambridge

This thesis is submitted for the degree of Doctor of Philosophy

January 2023

For my family, who deserve so much more.

“In order to learn, we must sometimes fall.”

Lelo Bokeria

DECLARATION

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text. It does not exceed 60,000 words, the prescribed word limit for the Clinical Medicine Degree Committee.

Work in Chapter 2 is a result of a collaboration with Rob Mok and Brett Roads. In particular, a large part of modelling work was done by Rob Mok and Brett Roads, with Brett Roads also performing the parameter estimations using the PsiZ model.

Work in Chapter 5 is a result of collaboration with Dingrong Guo, whose help in paradigm design and discussion of analysis strategies was invaluable.

The R codebase for simulating "power" for bayesian sequential designs that we have used throughout this thesis was designed by me in collaboration with Rik Henson, Alex Quent and Andrea Greve and is available publicly on the GitHub page of the MRC Cognition and Brain Sciences Unit.

Signed: 

Date: 12/01/2023

Levan Bokeria

Cambridge

ABSTRACT

One of the central challenges in cognitive neuroscience has been the study of internal mental representations of the external objects, events and relations that allow us to predict and interact with the world. Recently, researchers have uncovered parallels between the neural processing of physical space and of abstract knowledge, such that the established neural mechanisms for spatial navigation may also shed light on how we represent conceptual knowledge. In this thesis, we present a set of behavioural experiments examining the representational format of knowledge structures such as concepts and schemas, and develop learning paradigms that test algorithmic-level theories of spatial and non-spatial processing.

We start by discussing classical geometric models of knowledge representation, which view concepts as regions in abstract, multidimensional spaces organised by metric principles. These models have been supported by recent neuroimaging studies that suggest shared neural representations for spatial and non-spatial reasoning. We consider an older set of behavioural results that uncovered violations of the metric axioms of such representations, and discuss augmented geometric models that have been developed in response. One such model – the distance-density model – is examined in Chapter 2, using similarity judgments on a novel one-dimensional stimulus space. We did not find support for the basic prediction that psychological density affects similarity. In Chapter 3, we adapted the conceptual stimulus spaces used in the recent neuroimaging studies, and found that violations of metric requirements depend on the nature of the dimensions defining the stimuli. Nonetheless, using simulations and considering the prior psychological literature, we argue that another type of augmented model – the attention-weighted geometric model – is unlikely to account for such violations. These chapters therefore cast doubt on geometric models as adequate algorithmic-level theories for human knowledge representation.

The next two chapters develop schema learning tasks that lay the foundation for continued study of parallels between spatial and non-spatial reasoning. In Chapter 4, we examined how a non-spatial schema acquired in one conceptual space can influence learning in a different conceptual space. Across two experiments, we found effects consistent with generalisation of knowledge, but only for certain counterbalancing conditions. We argue for the importance of further refining our task and stimuli to develop a fast and flexible knowledge-transfer paradigm for studying relations between spatial and non-spatial

processing, which could also be extended to analogical reasoning, categorisation and schemas. In Chapter 5, we examined the nature of representational elements constituting spatial schemas. The prior literature has defined such schemas as networks of stimulus-location associative elements that can benefit learning. An unexamined possibility is that, instead of forming a cohesive network, such elements act independently to influence acquisition of new knowledge only within their local neighbourhood. Across two experiments involving learning of image-location associations on 2D boards, we find evidence consistent with this interpretation, and we outline how our paradigm can be adapted to address analogous questions for non-spatial schemas.

Taken together, our results question spatial representation of knowledge at the algorithmic level, as well as the nature of spatial schema, and emphasize the importance of continued research for elucidating commonalities and differences between spatial and non-spatial reasoning.

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Rik Henson who has been nothing short of an intellectual father for me during the last four years. He showed me what it is like to be a great supervisor and a friend at the same time. I know I drove him mad with my tendency to agonise over minor experimental details, constant misplacement of the article "the", subconscious love of Oxford commas, and requests to incorporate breaks during lab meetings. But hopefully, this has been balanced by me "letting" him beat me 6-0 in tennis, and perhaps some of the work in this thesis.

I would also like to acknowledge my lab "Rabble" and the wonderful friendly atmosphere it has provided. Thank you for the amazing time at work as well as outside, with punting trips, werewolf games and various shenanigans at academic conferences. I am proud to have been part of this team. Additionally, I am hugely thankful to the MRC Cognition and Brain Sciences Unit, which has been a home for last several years. I won't be the first or the last one to say that it is a magical place with no parallel anywhere in the world, with its coordinated coffee breaks, endless awkward corridors, tea-football games, Christmas pantos and, most importantly, a constant supply of interesting and fun people. Thank you to Rob Mok who has been a friend, a collaborator and an advisor, and a huge shoutout to the crazy girls from my office – the "89ers" Charlotte Garcia, Stepheni Uh and Tess Smith. Words cannot describe how glad I am to have been stuck with you.

My work also would not have been possible without my funder the Gates Cambridge Trust and my college Hughes Hall, with the financial, administrative, and emotional support they have provided. I have made many friends from Gates and Hughes Hall, and they have been an integral aspect of the "Cambridge experience."

Anything I have accomplished is a result of experiences along the way through various educational institutions I have attended and worked at. Thank you to the teachers, students and colleagues from the 42nd public high school back home, George Mason University, University of Rochester, Georgetown University, and the Donders Institute.

Finally, I would like to thank my friends and family back home, whose support and love transcended thousands of miles for years and have kept me going during the most difficult of times. This thesis is dedicated to my family, and especially to my grandfather Lelo who unfortunately did not make it to see me graduate… They deserve so much more, but I hope this work will count towards my effort to make them proud of me.

CONTENTS

LIST OF TABLES

List of Figures

List of Abbreviations and Acronyms

ATL	Anterior Temporal Lobe
ANOVA	Analysis of Variance
BF	Bayes Factor
EpCon	Episodes to Concepts
EHC	Entorhinal Cortex
fMRI	functional Magnetic Resonance Imaging
H1	alternative hypothesis
H0	null hypothesis
HPC	Hippocampus
ITI	inter-trial-interval
mPFC	medial Prefrontal Cortex
PA	paired-associate
MDS	Multidimensional Scaling
MDN	Multiple Demand Network
OFC	Orbito-Frontal Cortex

List of Appendices

1 INTRODUCTION

Humans have a remarkable capacity to learn about how the world works, and to represent this knowledge as rich internal models. We can extract statistical regularities which help us anticipate environmental phenomena, we categorize things and abstract them away into concepts which we communicate using complex compositional language, and we connect these concepts in meaningful relations that form hierarchies mirroring the complexity and causal dynamics of the real world. Such internal representations, in turn, have a top-down guiding influence on our subsequent learning and behaviour, impacting our perception, motor action, decision-making and formation of new memories. This ability to build flexible representations sets us apart from other animals, as well as state-of-the-art artificial intelligence, which still lacks the compositional dexterity and capacity to generalize acquired representations. It is not surprising therefore that attempts to understand the neural and computational bases of complex knowledge representation have been one of the central research areas in cognitive neuroscience. What exactly are concepts and how are they implemented in the human mind and brain? How and where are the relations between concepts represented, and how do such relational structures affect subsequent information processing and establishment of new internal structures?

When facing such daunting questions, a useful strategy is to break them down to multiple levels of abstraction. This was the approach taken by David Marr (1982) when he proposed three distinct but interrelated levels at which psychological processes could be analysed: *computational, algorithmic* and *implementational*. The computational level specifies the overall goal of a system. For example, Marr discussed how (one of) the primary goals of vision was to accurately detect shapes of objects and their arrangements,

to allow the organism appropriate interaction. At the algorithmic level, the transformation of the input-output of the system is described. We must specify the basic *representational primitives* of the system, and computations performed on them. For vision, Marr outlined the pipeline of transformation from 1D images, to 2.5D, to 3D internal representations giving shape and depth information. Finally, the implementational level specifies how the algorithm is instantiated in physical medium. Here, we would look at the organization and function of neural systems underlying visual processing.

Applied to the question of knowledge representation, research at each of Marr's three levels has a rich and deep history. Work presented in this thesis is most relevant to the algorithmic level, examining the format in which concepts and knowledge-structures such as schemas are represented, how they afford generalisation and influence learning. Specifically, we examine a recent proposal that spatial coding principles might underlie acquisition and organisation of non-spatial knowledge (e.g. Bellmund et al., 2018), and we develop experimental paradigms for efficiently studying relationships between spatial and non-spatial learning. In this introductory chapter, we start by discussing a prominent algorithmic-level theory of knowledge representation – geometric models of conceptual spaces. We continue by presenting how recent research at the implementational level has (at least indirectly) supported geometric models by demonstrating parallels between spatial and non-spatial neural coding. We further point out that an older set of behavioral studies from 1970s and 1980s have produced results incompatible with fundamental axioms of classical geometric models, which led to proposals of rival algorithmic-level representational theories: feature-based models. This tension motivates Chapters 2 and 3 of this thesis, where we use similarity judgment tasks to examine adherence of data to fundamental requirements of geometric theories. Following this, we present several open questions regarding generalisation of non-spatial schema knowledge. This sets the base for Chapter 4, where we present a new paradigm for systematically studying such generalisation in a controlled procedure and which can be adapted for examining transfer of schema knowledge between non-spatial and spatial domains. Finally, we discuss prior research on spatial schemas as networks of interrelated knowledge structures and present an important unanswered question facing the associated human and animal literature. Chapter 5 presents two experiments that attempted to shed some light on this challenge, and proposes future adaptations of such experiments for non-spatial schemas to further characterise the breadth of purported shared neurocomputational principles of spatial and non-spatial learning.

1.1 Geometric models of conceptual spaces

First algorithmic-level models of knowledge representation considered here are the geometric models (Balkenius & Gärdenfors, 2016; Carnap, 1928; Coombs, 1954; Gärdenfors, 2000; Markman, 2012; Shepard, 1958; Torgerson, 1965). Here, space is used as a representational medium for concepts and concept exemplars. A perceived object can be represented as a point with values along dimensions that correspond to its sensory or abstract qualities, such as length, brightness, or even political orientation. Thus, similar stimuli would have nearby positions, while dissimilar ones would be located further apart. If individual objects are points in a multidimensional space, then concepts can be defined as regions spanning multiple dimensions. Figure 1.1 illustrates this with a simplistic example, in a 2-dimensional "car space", where the dimensions specify weight and engine strength of a vehicle, and a concept of a *sports car* would be the bottom-right region of the space corresponding to low weight and a strong engine, while a concept of a *truck* would be in the upper right corner (Bellmund et al., 2018). Representation of a concept could additionally specify weighting of specific dimensions based on their salience as well as information about how different dimensions are correlated (Gärdenfors, 2000).

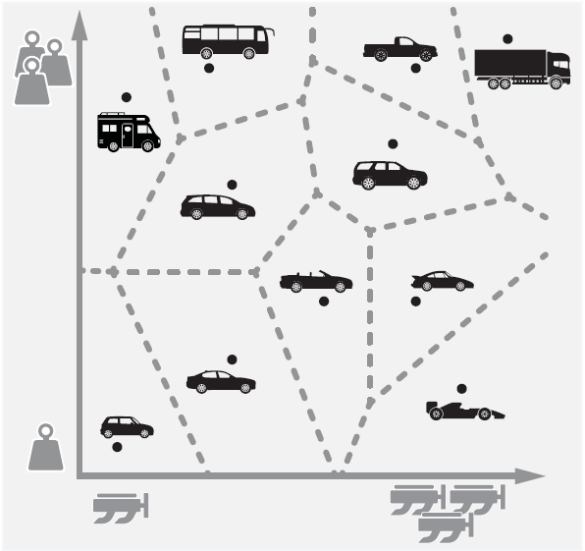


Figure 1.1: An example conceptual space.

A car space characterised by strength of the engine (*x* axis) and weight of the car (*y* axis). Each car is an exemplar with *x* and *y* coordinates, while a concept is a convex region in the space. For example, the concept of a *sports car* is defined by the bottom-right region of the space. Figure adapted from Bellmund et al. (2018). Reprinted with permission from AAAS.

In such a geometric knowledge representation, semantic similarity between concepts corresponds to the distance between them in the multidimensional space, and calculation of this distance often uses a *power metric* formula:

Equation 1.1: $$\boldsymbol{d(a,b) = \left[\sum_{i=1}^{N} |a_i - b_i|^{\gamma}\right]^{1/\gamma}}$$

where *d(a,b)* is the distance between points *a* and *b*, *i* corresponds to a specific dimension, *N* is the number of dimensions, and γ is the "Minkowski" distance metric that specifies how distances across dimensions are combined. When $\gamma = 2$, the metric becomes the familiar Euclidean distance, while when $\gamma = 1$, it corresponds to the *city-block* metric, where distances simply get summed across dimensions[1].

Behaviourally, distances between people's internal representation of concepts can be captured by asking them to rate similarities between them. A plethora of similarity judgment tasks have been used for this purpose (see section 1.6.4 below for some examples), and mathematical methods such as multidimensional scaling (MDS) were developed for using this similarity data to create visual depictions of judged items in lower-dimensional spaces (Borg & Groenen, 2005; Shepard, 1962; Torgerson, 1952, 1965). For example, Rips, Shoben and Smith (Rips et al., 1973) used MDS on people's similarity judgment of bird concepts to map-out a visual depiction of their internal psychological representations. The mapping showed that similar birds (like robin and sparrow) were located nearby, whereas others (like robin and goose) were further apart. Consistent with this idea of distances, the authors also found that people were faster to verify sentences like "A robin is a bird" than "A duck is a bird."

Dimensions can be of various types, affording different psychological processes to operate on them. For example, *psychologically separable* dimensions are those that can be attended to independently of each other, as for weight and engine strength in the car space example above (Garner, 1974; Maddox, 1992; Melara, 1992). Distances in such spaces are said to be characterised with a city-block metric, i.e. with the Minkowski γ parameter equal to 1 (see Equation 1.1). *Psychologically integral* dimensions are those that cannot be independently attended to, characterised instead with a Euclidean metric,

[1] Note that regardless of the specific metric that might underly the distance calculation (i.e. Euclidean metric, city-block metric, etc), the geometric models discussed in this thesis and the corresponding psychological literature are all Euclidean spaces, as opposed to spherical or hyperbolic spaces.

with $\gamma = 2$. For example, a colour can be decomposed into three different dimensions of hue, chroma and brightness, but people have difficulty judging these separately.

Geometric models gained extensive popularity in the second half of the 20th century, due to their elegance and efficiency in calculating distances using a simple procedure, and thanks to availability of techniques such as MDS, which seemed to visually capture intuitive correspondences of semantic similarity between various concepts. Extensive research was dedicated to employing similarity judgment tasks to recreate people's internal representation of various domains (e.g. Aisbett & Gibbon, 1994; Carroll & Arabie, 1980; Carroll & Wish, 1974; Coombs, 1954; Henley, 1969; Hutchinson & Lockhead, 1977; Kruskal, 1964b, 1964a; Monahan & Lockhead, 1977; Rips et al., 1973; Shepard, 1958, 1980; Torgerson, 1965).

1.2 Support for geometric theories from neural data

In Marr's framework, one way to assess the plausibility of algorithmic-level theories is to garner support from implementational-level findings, which can be more compatible with one algorithmic theory than another. Recent neuroimaging and physiological research have provided precisely such support for geometric theories, showing that the same neural systems involved when thinking and manipulating abstract concepts and knowledge structures are involved when people are navigating physical space (which is a special case of a geometric n-dimensional space, with n = 3). However, before expanding on this evidence, it is worth reviewing prior research on concept learning and storage in the brain.

Long-term conceptual representation has been extensively studied in in the field of semantic cognition (Rogers & McClelland, 2004). The *hub-and-spoke* model argues that modality-specific aspects of concepts are represented in brain regions responsible for processing the corresponding modality-specific information (i.e. *spokes*), while the anterior temporal lobe (ATL) functions as an integrator (i.e. a *hub*) of this distributed information (Lambon Ralph et al., 2017). Empirical support for this theory stems from computational, neuroimaging and lesion work, which have supported the central importance of the ATL as an integrating hub for multi-modal semantic cognition (e.g. Bozeat et al., 2000; Damasio et al., 1996; Hodges & Patterson, 2007; Jefferies et al., 2009; Lambon Ralph, 2014; Lambon Ralph et al., 2010; Lambon Ralph & Patterson, 2008; Snowden et al., 1989) . Other theories, however, emphasise the distributed nature of semantic cognition without a need for a centralising hub region, arguing that different

conceptual domains are represented in anatomically distinct and functionally independent regions in the ventral visual pathway (e.g. Capitani et al., 2003; Chouinard & Goodale, 2009; Kanwisher, 2010; Mahon et al., 2009).

In terms of concept acquisition, one of the crucial brain regions has been the hippocampus (HPC). The Episodes-to-Concepts (EpCon) model proposed by Mack et al. (2018) outlines various component processes necessary for concept formation, such as biasing attention to specific features, pattern separation as well as pattern completion, sensitivity to prediction error, and integration of different item representations. Mack and colleagues argued that the HPC (in coordination with prefrontal cortical regions) is perfectly suited to support these functions, and reviewed neural evidence of HPC involvement during early concept learning (e.g. Davis et al., 2012b, 2012a, 2014; Kumaran et al., 2009; Mack et al., 2016, 2018; Schapiro et al., 2012).

For a long time, the HPC was not considered to be involved in representation of longer-term semanticized conceptual knowledge. For example, patients with damage to HPC do not normally show deficits in semantic cognition (Blumenthal et al., 2017; Knowlton & Squire, 1993; Squire & Knowlton, 1995). Much work in human cognitive neuroscience instead outlined the importance of HPC for episodic declarative memory (Cohen & Squire, 1980) and general relational reasoning (Cohen & Eichenbaum, 1993) that supports forming associations among items and context to bind them into a coherent event representation that can be consolidated into memory (Eichenbaum & Cohen, 2004; Knierim et al., 2014). However, in recent years, HPC and its neighbouring regions (particularly the entorhinal cortex, EHC) have become centre stage candidates as hubs for long-term conceptual processing as well. This has emerged from continued attempts to reconcile the role of HPC in general cognition (as outlined above) and its extensive involvement in supporting spatial navigation (Buzsáki & Moser, 2013; Eichenbaum & Cohen, 2014).

The HPC is home to *place cells*, which have been found to track an animal's location within an environment (O'Keefe John, 1978), and is thought to create an allocentric *cognitive map* of an environment (although see Eichenbaum & Cohen, 2014 for a challenge to this view). The EHC, on the other hand, contains *grid cells*, which tile the navigable space in equilateral triangles, firing at systematic intervals (Hafting et al., 2005). Grid cells are thought to provide a metric for calculating distances and vector relationships in a 2D space (McNaughton et al., 2006). Grid-like activity has also been found in humans navigating virtual physical spaces (Doeller et al., 2010), in terms of the

six-fold modulation of the fMRI signal (Figure 1.2-A) that would be predicted by the hexagonal organisation of grid cells. Interestingly, this study showed that grid-like activity was observed not only in the entorhinal cortex but other regions too, notably the medial prefrontal cortex (mPFC), where later intracranial recordings would confirm existence of grid cells (Jacobs et al., 2013).

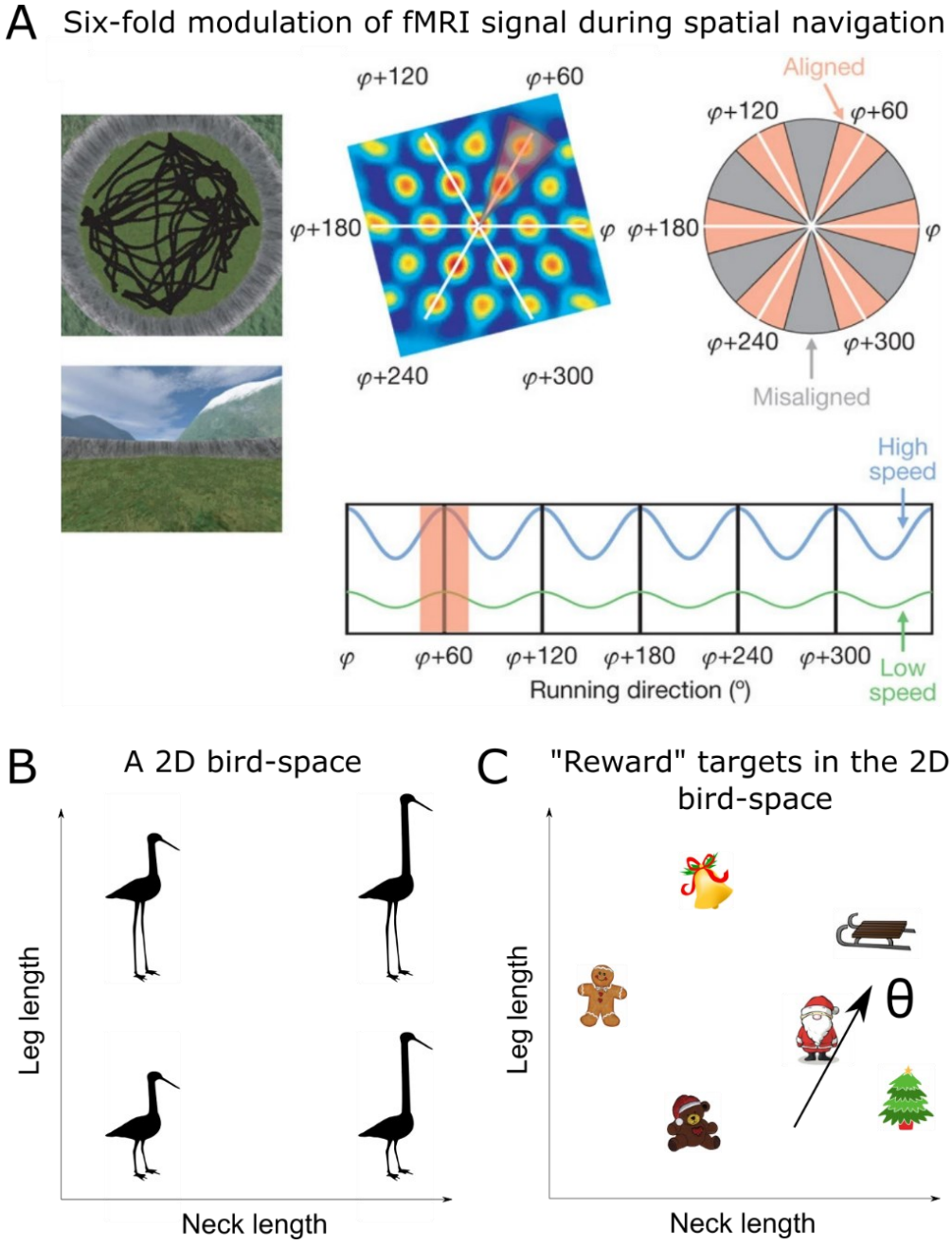


Figure 1.2: fMRI studies of navigation in physical and conceptual spaces.
(A) Doeller et al. (2010) had participants navigate a circular arena. Left two panels display navigation trajectories (top) and the participant view (bottom). The middle heat plot shows an autocorrelogram of a typical grid cell, with the three main axis of the grid (white lines). The top-right panel schematically depicts running

orientations that are either aligned (red) or misaligned (grey) with the grid axes. The bottom panel shows predicted sinusoidal modulation of fMRI signal depending on the running direction, with the running speed determining the strength of the effect. (B) The two-dimensional neck:legs space of Constantinescu et al. (2016). (C) Distribution of "reward" Christmas toys associated with specific exemplars in the bird space of Constantinescu et al. (2016). "Movement" (i.e. morphing of the birds with a certain neck:legs ratio) at a particular angle θ would be either aligned or misaligned with grid axes, allowing for a test for six-fold modulation of the fMRI signal. Panel (A) adapted from "Evidence for grid cells in a human memory network", Doeller, C. F., Barry, C., & Burgess, N., Nature, Volume 463, 2010. The Licensed Material is not part of the governing OA license but has been reproduced with permission from SNCSC. Panel (C) adapted from Constantinescu et al. (2016). Reprinted with permission from AAAS.

In 2016, Constantinescu and colleagues found similar hexadirectionally modulated fMRI signal coming from the same brain regions (EHC, mPFC and PPC, among others) when participants "navigated" a conceptual space. The authors exposed participants to a two-dimensional *bird space*, where exemplar birds varied along the dimensions of neck length and leg length (Figure 1.2-B). Across several training days, participants learned to morph specific exemplars into other exemplars by smoothly changing the neck and leg length. Within the 2D space, specific bird exemplars were associated with arbitrary "rewards" (Christmas toys) which participants "discovered" as they "navigated" through the 2D space. Crucially, any specific morphing was associated with movement in the 2D neck-legs space at a specific angle, analogous to navigation in a physical space at a specific angle. This allowed authors to systematically look for brain regions where activity was hexadirectionally modulated, indicative of an underlying grid-like neural activity.

Other studies have focused on the role of the HPC in representing distances in multi-dimensional conceptual spaces. Theves et al. (2019) taught participants a 2D stimulus space containing artificial stimuli consisting of a square and a circle that varied along the dimensions of the opacity of the square and the size of the circle (Figure 1.3). The participants learned that specific object exemplars were associated with certain images (houses, furniture and everyday items). Using the degree of repetition-related suppression of fMRI data (when features are repeatedly processed, Grill-Spector et al. 2006), as well as representational similarity analysis on multivoxel representational patterns

(Kriegeskorte, 2008), the authors showed that HPC tracked Euclidean distance between the associated images in the underlying 2D space.

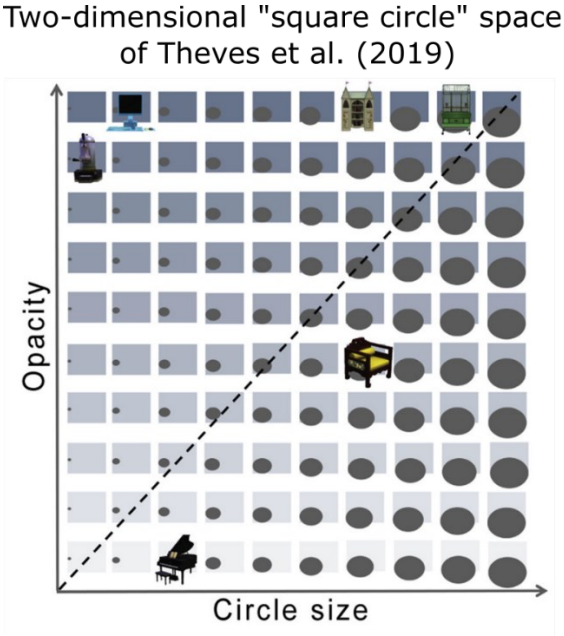


Figure 1.3: The two-dimensional feature space of Theves et al. (2019). Participants learned associations between specific exemplars from this abstract stimulus space and certain images of items (furniture, buildings, etc.). Figure used with permission of Elsevier Science & Technology Journals, from "The Hippocampus Encodes Distances in Multidimensional Feature Space", Theves S., Fernandez, G., & Doeller, C. F., 29(7), 2019, 1226-1231.e3, year of copyright 2023; permission conveyed through Copyright Clearance Center, Inc.

To date, such parallels between navigation-like neural activity underlying manipulation of conceptual spaces have been shown in many other paradigms involving various types of dimensions, such as social spaces (Park et al., 2020; Tavares et al., 2015), odour spaces (Bao et al., 2019), value spaces (Knudsen & Wallis, 2021), multi-modal visual-auditory spaces (Viganò & Piazza, 2020) and others (Theves et al., 2020). Theorists speculated that perhaps the evolutionarily old neural machinery used for spatial navigation got "reused" for navigating higher-dimensional knowledge structures such as conceptual spaces (Bellmund et al., 2018; Buzsáki & Moser, 2013). Proposals emerged that viewed the hippocampal-entorhinal system as a hub for concept processing, enabling manipulation of conceptual knowledge at multiple scales (Bellmund et al., 2018; Morton & Preston, 2021). This echoed Tolman's view of parallels between spatial maps and more abstract cognitive maps, whereby animals form rich internal models of the world consisting of knowledge structures extracted from experience (Tolman, 1948).

So far, we have reviewed evidence in support for geometric theories at behavioural and neural levels. In the following sections, we outline an earlier line of behavioural evidence that challenged geometric theories but which has been largely ignored in the recent excitement over parallels between spatial and non-spatial processing (Bokeria et al., 2021). This motivated Chapter 3 of this thesis, where we adapted stimuli from Constantinescu et al. (2016) and Theves et al. (2019), and used similarity judgment tasks to test adherence of such 2D spaces to certain fundamental requirements of geometric models.

Prior to reviewing these algorithmic-level challenges, it is important to note here that the proposal that neural computations for spatial processing have been adapted for conceptual processing has also been challenged at the implementational level. For example, Mok & Love (2019) argued that instead of spatial navigation machinery being adapted for conceptual reasoning, both processes are driven by a domain-general learning mechanism centred on a clustering algorithm. Clustering models can capture learning principles and representational nature of conceptual spaces (Davis et al., 2012a; Love et al., 2004; Mack et al., 2016), where exemplars are often clumped at particular locations and do not span the entirety of the feature space. During physical navigation, however, an animal will often explore the full range of the available space. Using simulations, these authors showed how exploration of physical space leads to an ordered set of clusters that resemble grid-cell activity, proposing that perhaps grid-cell firing reflects the error monitoring output of the learning algorithm instead of provision of a metric signal that allows representation of animal's position. Still, other proposals have argued that hippocampal-entorhinal system creates topological representations which reflect temporal contiguities between experiences (Rueckemann et al., 2021). In physical space, temporal contiguity is highly correlated with physical proximity. These authors argued how this characterization better explains a plethora of results showing HPC-EHC involvement in various spatial and non-spatial tasks.

1.3 Challenges to geometric models and feature-based representations

1.3.1 Axioms of geometric models

In the midst of their popularity during the mid-20th century, behavioural evidence started to emerge against the validity of geometric models as appropriate algorithmic-level descriptions of internal psychological representations. For example, Gati and Tversky

(1982) showed that similarity between a pair of perceptual or conceptual stimuli increases as a result of addition of the same feature to both items. If items are points in a multidimensional space and dissimilarity is a metric distance between them, then addition of another dimension along which both items have the same coordinate should not change the outcome of distance calculation in Equation 1.1. Tversky and Gati (1982) further argued that similarity data purportedly generated by underlying geometric representation of items consistently violated various axiomatic requirements of geometric models. In a series of theoretical and empirical papers, they and other theorists presented detailed analysis of necessary and sufficient requirements for geometric models (Beals et al., 1968; Beals & Krantz, 1967; Burns et al., 1978; Gati & Tversky, 1982; Krantz & Tversky, 1975; Tversky & Gati, 1982; Tversky & Krantz, 1969, 1970; Wender, 1971; Wiener-Ehrlich, 1978). These requirements were:

- Minimality: $d(a,b) > d(a,a) = 0$. The distance between two points must be larger than the distance between a point and itself.
- Symmetry: $d(a,b) = d(b,a)$. The distance between points *a* and *b* must equal the distance between *b* and *a*.
- Triangle inequality: $d(a,c) \leq d(a,b) + d(b,c)$. The shortest path between two points must be a direct line; not a path going through a third, outside point.
- Segmental additivity: $d(a,c) = d(a,b) + d(b,c)$. For any three points lying on a straight path, distances along each segment of that path must be additive.

Violations of these requirements have been extensively documented in literature. For the minimality assumption, studies have shown that measures of self-similarity are not constant across different items (e.g. Rothkopf, 1957), and researchers have noted that the off-diagonal entries in a matrix of pairwise similarity values sometimes exceed diagonal ones, meaning that two different items are judged more similar than an item to itself (discussed in Krumhansl, 1978 and Tversky, 1977).

Violations of the symmetry requirement have been found using similarity or dissimilarity judgments (Tversky, 1977), identification confusion tasks (Appelman & Mayzner, 1982; Gilmore et al., 1979; Keren & Baggen, 1981; Townsend, 1971), or discrimination confusion tasks (Rothkopf, 1957). Tversky (1977) elaborated that similarity judgments often take a directional form such as "assess the degree to which *a* is similar to *b*", where *a* is taken as the subject while *b* is the referent. In different experiments using either countries varying in prominence, or geometric figures varying in "goodness of form", he

showed that swapping the subject-referent in the directional similarity statements resulted in large asymmetries. For example, North Korea was rated to be more similar to Red China than Red China to North Korea, and irregular geometric forms were more similar to "good forms" than vice versa. Tversky assigned this effect to the differential salience of the geometric forms. Relatedly, Rosch (1975) argued that in various perceptual or semantic categories such as colours, line orientations, or numbers, certain prototypical stimuli serve as cognitive "reference points", such as the colour red, a perfect square, or multiples of 10. As a result, non-prototypical stimuli are rated as more similar to prototypes than the other way around.

Tests for the triangle inequality and segmental additivity are trickier since they require similarity or dissimilarity measures on a continuous (interval or ratio) scale, while most tasks only provide ordinal data where rounding continuous psychological similarities to a discrete scale can result in distortions. To circumvent this limitation, Tversky and Gati (1982) developed an ingenious method to test for the triangle inequality using only ordinal measures of similarity (discussed in Chapter 3). While reviewing prior data as well as presenting new experimental results, the authors showed systematic violations of the triangle inequality during dissimilarity and similarity judgments of stimuli from various 2-dimensional spaces.

1.3.2 Feature-based models

In response to these violations of axioms of geometric models, Tversky developed an alternative algorithmic-level theory of knowledge representation and similarity computation based on feature sets (Tversky, 1977). Concepts or exemplars consist of discrete sets of features, and a similarity comparison involves contrasting shared and unique features. Taking example items *a* and *b* with their associated feature sets denoted by *A* and *B*, the *contrast model* is formulated as follows:

Equation 1.2: $$\boldsymbol{S(A,B) = \theta \times f(A \cap B) - \alpha \times f(A - B) - \beta \times f(B - A)}$$

where $(A \cap B)$ represents shared features of *a* and *b*, $(A - B)$ represents features unique to *a* and $(B - A)$ represents features unique to *b*. θ, α, and β are weights for the common and distinctive feature sets, while *S* represents the similarity. The scale *f* reflects the prominence of various features, and thus measures the relative contribution of particular features to the similarity. The scale value *f(A)* for a stimulus *a* is taken to represent overall salience of the stimulus.

Representation of stimuli as feature sets had been widely employed in characterization of different cognitive processes, such as perceptual learning (Gibson, 1969), speech perception (Blumstein & Stevens, 1981; Jakobson et al., 1961), semantic judgment (Smith et al., 1974), or preferential choice (Tversky, 1972).

The contrast model does not predict that self-similarity will be equal for all stimuli. For identical items, the first component of the model $\theta f(A \cap A)$ will be the sole determinant of the resulting self-similarity $S(A, A)$, and will be proportional to the richness of the set of features comprising the stimulus. Certain self-similarities could be smaller than similarities between different pairs of items too, if the feature-set overlap of those different items is large and they do not have too many unique features[2].

Explaining asymmetries requires an additional assumption of the "focusing hypothesis" (Tversky, 1977). In directional similarity questions such as "assess the degree to which *a* is similar to *b*", there is naturally a larger focus on the subject *a* compared to referent *b*. In terms of the contrast model, this is equivalent to $\alpha > \beta$, which in turn means that the distinctive features of *a*, *(A – B)*, will contribute more to the reduction of similarity than distinctive features of *b*, *(B – A)*. If the referent of the similarity comparison statement is a prototypical concept, or one that is high in saliency (i.e. high in its number of features), then the resulting total similarity will be greater than when a less salient item is the referent and the prototype is the subject. This would readily explain results such as North Korea being rated more similar to Red China than the other way around.

In their 1982 paper, Tversky and Gati presented a detailed set-theoretical analysis of how the contrast model coupled with additional assumptions on additivity of features within and across properties, can produce violations of the triangle inequality.

1.3.3 Augmented geometric models

1.3.3.1 The distance-density model

Instead of dismissing geometric models, some researchers have proposed modifications to account for violations of their axioms. Summarising this line of reasoning, Nosofsky (1992) discussed that many cognitive tasks such as similarity computations should be

[2] Although the contrast model can explain minimality violations when self-similarities are smaller than similarities between other pairs, i.e. when S(A,A) < S(B,C), it cannot explain how self-similarity can be smaller than the similarity of that same item with a different item, i.e. S(A,A) < S(A,B)

viewed as *representation-process* pairs, whereby items receive certain internal psychological representations in mind while various cognitive processes then act upon them, which importantly vary with the task at hand. A general form of such a process model, applied to similarity judgments, states that proximity of items *a* and *b* is given by

Equation 1.3: $$\boldsymbol{p(a,b) = F\big(s(a,b) + r(a) + c(b)\big)}$$

where *s(a,b)* is the symmetric similarity, while *r(a)* and *c(b)* are bias terms associated with each item (Holman, 1979). The symmetric similarity measure *s(a,b)* in the above formula could be a result of feature-comparison, as in the first component of Tversky's contrast model, or a result of distance calculation in a multi-dimensional geometric space, as postulated by classical geometric models. Taking the latter approach, Krumhansl (1978) proposed the *distance-density* model (as a special case of Nosofsky's framework), where the dissimilarity between items *a* and *b* are a result of some "objective" metric distance between the points, as well as the local density around *a* and *b*:

Equation 1.4: $$\boldsymbol{d'(a,b) = d(a,b) + \alpha \times D(a) + \beta \times D(b)}$$

This model suggests that dense regions of the stimulus space are expanded, resulting in stretching of the psychological space and large distances between points. Coupled with the focusing hypothesis, the distance-density model readily explains asymmetric judgments. If $\alpha > \beta$ during directional similarity judgment, then *d'(a,b) > d'(b,a)* whenever *D(a) > D(b)*. Krumhansl argued that prototypes and salient stimuli are typically in denser regions of the space, explaining the asymmetric judgments. This model could also explain violations of minimality, since if $\alpha = 1$ and $\beta = 1$, then *d'(a,a) = d(a,a)+D(a)+D(a)*, which could be bigger than *d'(a,b) = d(a,b)+D(a)+D(b)* if *D(a)* much bigger than *D(b)*, and *d(a,a)* not much smaller than *d(a,b)*.

In Chapter 2, we designed an experiment specifically geared towards testing the basic prediction of the distance-density model, namely that similarities between items should increase once density is increased in their neighbouring regions. In brief, we did not find support for the distance-density model, arguing that it might not be a sufficient answer to account for the axiomatic violations outlined by Tversky.

1.3.3.2 Attention-weighted geometric models

A further strategy for augmenting geometric models is to incorporate attentional processes that can selectively change salience of certain dimensions depending on the context (Nosofsky, 1986; Smith & Heise, 1992). Gärdenfors' recent resuscitation of

geometric models (Gärdenfors, 2000) explicitly acknowledges the context dependent nature of similarity comparisons, and presents a modified version of the Equation 1.1 distance calculation formula:

Equation 1.5: $$\boldsymbol{d(a,b) = \left[\sum_{i=1}^{N} w_i \times |a_i - b_i|^{\gamma}\right]^{1/\gamma}}$$

where w_i is the attention-weight given to dimension i. Thus, depending on a particular context, only certain dimensions become relevant through dynamic attentional weighting of pertinent features. Although Gärdenfors does not systematically flesh out the specific conditions directing such changes in attention, such modified geometric model could in principle account for asymmetric similarity judgments (Decock & Douven, 2011). In directional similarity statements of the form "how similar is *a* to *b*", the dimensions might be weighted differently relative to the statement "how similar is *b* to *a*".

The process of feature selection and the role of such attentional processes has been long acknowledged by researchers studying similarity. For example, Sjöberg argued that similarity comparisons involve an active search of properties along which items are similar (Sjöberg, 1972). Tversky himself mentioned that: "When faced with a particular task (e.g., identification or similarity assessment), we extract and compile from our data base a limited list of relevant features on the basis of which we perform the required task. Thus, the representation of an object as a collection of features is viewed as a product of a prior process of extraction and compilation." (Tversky, 1977).

Importantly, attentional shifts in the features/dimensions of comparison could account for violations of the triangle inequality as well. Consider a famous example discussed by Willian James (1890), whereby people judge the moon to be similar to a gas jet, but also similar to a football. However, a football and a gas jet are not considered to be similar at all, leading to a violation of the triangle inequality. However, people likely shift criteria between these pairs, using luminosity to compare the moon and a gas jet, but using shape to compare the moon and a football. Tversky and Gati (1982) acknowledged that such systematic shifts in reference frame could account for violations of the triangle inequality in their own studies. This would require, however, that the set of attention weights are not only context specific, but trial specific. They provided analytical and theoretical reasons why such an attentional account would not be an appealing explanation for their data.

In Chapter 3, we adapted various 2D stimulus spaces that have been used in recent neuroimaging literature to test their adherence to requirements of segmental additivity and the triangle inequality. This also provided an opportunity to check the validity of

attention-weighted geometric models. In brief, our data could not exclude the possibility that some of the violations in our stimulus spaces could be accounted for by the attention-weighted geometric model. However, in combination with prior findings in perceptual processing literature, we argue that the likely explanation behind our axiomatic violations is the non-linear mapping between physical and psychological distances, which cannot be accounted for through incorporation of attentional processes.

1.4 Generalisation of schema knowledge across conceptual spaces

Apart from studying the format in which individual conceptual knowledge is represented, another crucial question concerns how such concepts interact with each other. Related to this, while discussing parallels between spatial and non-spatial processing and proposing a role of HPC-EHC system for organizing knowledge across multiple domains, Bellmund and colleagues (2018) outlined that one of the pressing questions concerns how information encoded in distinct domains can interact. Can knowledge acquired in one abstract space be transferred to another to facilitate (or perhaps inhibit) learning? What is the neural basis of such transfer? If spatial and non-spatial reasoning share similar neural coding principles, would the dynamics of knowledge transfer between conceptual spaces be similar to that between conceptual and physical spaces? In Chapter 4 of this thesis, we designed a knowledge transfer paradigm, adapting the simple 2D bird space used by Constantinescu et al. (2016).

In the learning task of Constantinescu and colleagues (2016), the associations between specific bird exemplars and Christmas toy rewards can be viewed as landmarks in the 2D neck-legs space. Just like in a physical space, these landmarks form a geometric structure with a specific shape, representing a non-spatial associative knowledge or a *schema*. Schemas have historically been characterised as structures that can guide the interpretation of new events and aid in encoding and retrieval of new memories (Bartlett, 1932; Piaget, 1926, 1952; Tulving, 1972). We asked whether learning such a non-spatial schema of landmark arrangements in one bird space could aid learning in a different bird space with different dimensions, but where the landmarks were arranged in a similar shape. In such a paradigm, various factors can be systematically examined for their influence on knowledge transfer. For example, how does similarity between the defining dimensions of two stimulus spaces impact successful generalisation? Does degree of underlying structural similarity in landmark arrangements parametrically modulate amount of knowledge transfer, or is generalization an all or none phenomenon?

Furthermore, does a more abstract schema develop if one encounters multiple 2D spaces with similarly arranged landmarks, and how would generalisation enabled by such higher-order structure be different from generalization afforded by more “concrete” schemas instantiated with specific sensory stimuli?

Designing an efficient and flexible paradigm to answer such questions would benefit multiple lines of research studying the psychological processes involved in extraction of structure and inference during problem solving, and the neural basis of such generalisation (Taylor et al., 2021). For example, in the field of analogical reasoning (Holyoak, 2012), researchers have long studied processes underlying transfer of knowledge from one domain to another (e.g. Catrambone et al., 2006; Gick & Holyoak, 1980, 1983; Holyoak & Koh, 1987). It is thought that such generalization relies on structured representations in the two domains, whereby elements are analysed in terms of their role-based relational properties (Gentner & Markman, 1997). Figure 1.4 depicts a schematic representation of how such analogical reasoning progresses. A target domain cues a retrieval process for a source domain, after which a mapping is accomplished between relational elements, allowing transfer of existing knowledge. For example, consider someone learning about the structure of an atom, who might benefit from knowing how the solar system is organized. Once the source knowledge is retrieved, a mapping is established in which the solar system can be compared to an atom when analysed in terms of “role”-“filler” relationships: The sun and nucleus act as “fillers” for the “role” of “being at the centre”, while planets and electrons “fill” the “role” of revolving objects. Transfer of knowledge might happen if one knows that the sun is much heavier than the revolving planets, allowing inference that the nucleus is heavier its revolving electrons. In time, multiple exposures to similar role-filler relationships induces an abstract schema devoid of specific instances, which acts as an independent cognitive structure for interpreting new incoming information (Holyoak, 2012).

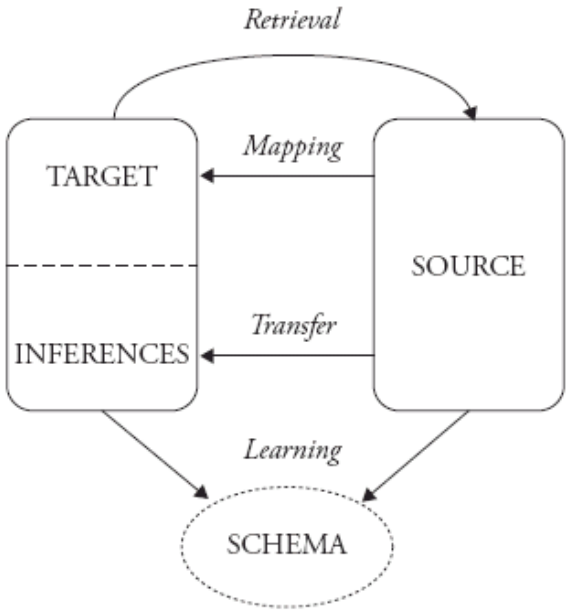


Figure 1.4: Central components of analogical reasoning.
Once a participant encounters a target domain to which generalisation must occur, a retrieval process finds a suitable source domain knowledge. A mapping is established between the elements of the source and the target domains, and a transfer of knowledge might occur to enable making novel inferences. Over time, with repeated analogical transfer, a more abstract schema might develop that acts independently to support generalisation. Figure used with permission of Oxford University Press - Books (US & UK), from "Analogy and Relational Reasoning", Holyoak, K. J., 2012, year of copyright 2023; permission conveyed through Copyright Clearance Center, Inc.

In problem solving research, some studies have found analogical transfer of solution strategies to be extremely rare, often requiring of an explicit hint (Gick & Holyoak, 1980). Others, on the other hand, have reported spontaneous implicit generalisation between problems with shared structure but very different surface qualities (Day & Goldstone, 2011). What determines the success of such structural alignment is still an open question. In her structure-mapping theory, Gentner (1983) argued that a "good" analogy involves alignments with high degree of structural parallelism as well as systematicity. Holyoak and colleagues expanded this view in their *multiconstraint theory* (Holyoak & Thagard, 1989), which postulated that a coherent analogy requires alignment at multiple levels – surface similarity, structural relations and functional properties of involved elements. Our attempt in Chapter 4 was to design a paradigm that would allow examination of systematic effects of such convergence or divergence of alignment at surface or structural

levels. Furthermore, as we are unaware of any published studies examining analogical transfer across spatial and non-spatial knowledge domains, our setup for studying generalisation across conceptual spaces could be adapted to study conceptual-to-physical knowledge transfer and test the applicability of proposals such as the multiconstraint theory.

1.5 Nature of spatial schemas and their role in knowledge acquisition

Apart from their role in supporting knowledge transfer across domains, schemas have been shown to influence acceleration of within-domain learning (e.g. Tse et al., 2007, 2011; van Buuren et al., 2014; van Kesteren et al., 2010, 2013; Wang et al., 2012; for reviews see van Kesteren et al., 2012; Fernández & Morris, 2018; Ghosh & Gilboa, 2014; Gilboa & Marlatte, 2017). Specifically, incoming information consistent or *congruent* with an existing schema is learned and retained better than inconsistent information (e.g. van Kesteren et al., 2020). In the last chapter of my thesis, we examined the representational format of spatial schemas and processes that underlie such schema-based acceleration of knowledge integration, and suggest extension of such paradigms for the study of analogous questions on non-spatial schemas.

Systematic investigation of the neural basis of spatial schemas and their influence on learning was initiated when Tse and colleagues (2007) created a rodent schema task. In their study, rats underwent extensive training to learn certain flavour-place associations in an environment, whereby a certain flavour in the start box predicted existence of food with same flavour in a particular location in the environment. As training progressed, the consistent flavour-place paired-associates (or PAs) formed a stable spatial schema in the form of a network of knowledge structure (similar to the landmarks proposed in the earlier example of generalisation across conceptual bird spaces). Crucially, when rats had to learn a new PA within the same environment, learning was accelerated thanks to the existing schema, and the new PA knowledge became hippocampus-independent within 48h, much faster than normal systems consolidation of new memories. Later studies demonstrated the importance of medial prefrontal cortex (mPFC) in schema-based learning and recall of such memories (Tse et al., 2011; Wang et al., 2012), indicating that, once consolidated, schema-related memories rely on prefrontal cortical structures as opposed to the HPC.

Such paired-associate learning tasks have been adapted for humans to facilitate comparison of processes across species (e.g. Guo & Yang, 2020, 2022; Schott et al., 2019; Sommer, 2017; van Buuren et al., 2014). For example, van Buuren and colleagues (2014) trained their participants across multiple days to learn picture-location associations on a 2-dimensional board displayed on a computer screen. These picture-location PAs formed a stable schema, which was shown to aid in subsequent learning and retrieval of new PAs. In a similar paradigm, Guo and Yang (2020) trained participants on picture-location associations on 2D boards, and again showed that having an existing network of such associations established over multiple days of training aids the learning of new picture-location associations.

One outstanding question concerns the precise mechanism by which learning of new knowledge is accelerated in paradigms reviewed above. It is possible that when learning any particular new picture-location association, the learning is accelerated not due to the whole network of existing PA structure (i.e. the schema), but only thanks to the most proximal schema components, which act as isolated landmarks onto which the new knowledge can be scaffolded. In other words, perhaps during the initial training phase, instead of forming a network of knowledge items, each PA is encoded and consolidated as a distinct element, aiding in subsequent acquisition of those new PAs that happen to be close-by. In the previous paradigms discussed above, it was not possible to differentiate between such "local" vs "global" effects of schemas, since every new learned PA was directly adjacent to an old PA. Chapter 5 describes a new spatial paired-associates learning task, where we tested whether these knowledge structures have a global facilitatory influence due to their interconnected network-like nature, or whether each element is encoded separately, and only locally helps learning of neighbouring new paired-associates.

1.6 Using similarity to study cognition

Chapters 2 and 3 of this thesis use various similarity judgment tasks to examine theories about knowledge representation. Chapter 4 focuses on generalisation of knowledge, which is also thought to rely on a higher-order notion of similarity. Therefore, this section presents a brief history of similarity as a subject of study in cognitive science, including various types of tasks that have been used to measure similarity.

1.6.1 Popularity of similarity

Similarities normally reflect an inherent order in the world, whereby similar items tend to behave similarly. As Quine (1969) stated: "Similarity is fundamental for learning, knowledge and thought, for only our sense of similarity allows us to order things into kinds so that these can function as stimulus meanings. Reasonable expectation depends on the similarity of circumstances and on our tendency to expect that similar causes will have similar effects (p. 114)." William James (1890) argued that "This sense of Sameness is the very keel and backbone of our thinking". As a psychological construct, similarity has permeated the study of many facets of cognition, such as problem solving, attention, prediction, memory, categorisation and perception (Goldstone & Son, 2012). For example, transfer of learned responses depends on the similarity of the situation at-hand and the original training context (Osgood, 1949; Thorndike, 1931), and remembering is influenced by the similarity between the encoding and retrieval environments (Roediger, 1990). If an event triggers a memory of a similar event in the past, the memory might guide predictions in the present environment (Sloman, 1993; Tenenbaum & Griffiths, 2001), while categorization of an unknown new object has been shown to depend on its similarity to known objects (Nosofsky, 1986). Similarity has also been used as a tool for characterising the structure of cognitive entities and processes; for example, experts and novices can be differentiated based on the depth at which they see similarities between two situations or problem sets (e.g. Hardiman et al., 1989; see Sjöberg, 1972 for other examples).

1.6.2 Critics of similarity

The notion of similarity has not been without its serious critics. Various theorists (Medin et al., 1993; Murphy & Medin, 1985) have emphasized that similarity is perhaps too flexible as a concept, and too dependent on context and task. Goodman's (1972) critique was especially sharp, pointing out that statements like "A is similar to B" are ill-defined and require a specification of "with respect to property Z". But once such specification is provided, the notion of similarity becomes superfluous, with all explanatory work being done by the "with respect to…" statement. He further fleshed out the problem of context dependency, pointing out that for any given situation, only a subset of features belonging to the object are relevant, and even within this subset, not all features are weighted uniformly. More recently, Medin and colleagues (1993) have presented experimental data arguing that similarity judgments are not only task and context dependent, but also shaped

by the specific pair of items being compared, with different features being emphasized on different trials. This seemingly supported Goodman's verdict for similarity, that it was "invidious, insidious, a pretender, an imposter, a quack" (Goodman, 1972, p. 437).

1.6.3 In defence of similarity

Throughout the years after Goodman's critique, theorists and experimentalists started building a case to defend the concept of similarity. First, the process whereby various activated features are combined to produce an overall judgment of similarity is indeed complex, but can still vary systematically across different domains or tasks (e.g. Goldstone & Son, 2012; Medin et al., 1993; Tversky, 1977). Therefore, Goodman's point was not completely fair when arguing that specification of "with respect to property Z" renders "similarity" as vacuous: There is more to similarity than just specification of relevant features – such as the important psychological processes of property combination and integration – and these should be studied and characterised (Medin & Schaffer, 1978; Nosofsky, 1992a). Second, as has already been discussed above, context-dependency of similarity judgments has been explicitly acknowledged by many theorists (Goldstone et al., 1997; Medin et al., 1993; Nosofsky, 1986, 1992b; Sjöberg, 1972; Tversky, 1977). Decock and Douven (2011) have argued that both Tversky's feature-matching model (Equation 1.2) and Gärdenfors' incorporation of attentional weighting in geometric conceptual spaces (Equation 1.5) can account for contextual variation in similarity judgments. Even though Medin and colleagues showed that similarities vary even on the particular pair of objects at hand, they argued that such dependence is systematic and should be the proper focus within the psychological study of similarity, stating that "our thesis is that there are systematic and well-structured patterns to how multiple pieces of information are structured to yield similarity assessments" (Medin et al., 1993, p.258). Thus, the construct of similarity as a psychological process and as a subject of study has been repeatedly defended, and has continued to be used for uncovering structures of mental representations (e.g. Hebart et al., 2020; Love & Roads, 2021), while acknowledging the daunting challenges of systematising and characterising the underlying task and context dependent representations and processes.

1.6.4 Brief overview of similarity judgment tasks

Having presented a case for why similarity is a worthwhile construct for study, we next present a brief overview of popular similarity judgment tasks that have been used in the literature, outlining their pros and cons. Throughout various experiments described in this

thesis, the choice of these similarity tasks has been guided by theoretical and practical considerations which are explained in the appropriate chapters.

1.6.4.1 Pair-wise similarity judgment task

During a pair-wise judgment task (e.g. Krantz & Tversky, 1975), pairs of stimuli are presented to participants along with a Likert-style rating scale (e.g. 1 through 9) to indicate perceived similarity or dissimilarity. An advantage of this method is that each pair is viewed in isolation, and no constraint is placed on the comparison process (Kriegeskorte & Mur, 2012). Thus, pair-wise ratings can capture multidimensional dissimilarity relationships. However, this can also be viewed as a con, because although no external constrains are provided, unobserved internal trial-by-trial shifts in criteria might develop throughout the experiment (per suggestion of Medin et al. 1993). This likely contributes to the relative noisiness and inconsistency of this method compared to others (Demiralp et al., 2014; Li et al., 2016). Furthermore, if the scale does not offer enough range to capture fine-grained differences in perceived similarities (e.g., when using a limited 5-point scale), this could result in a loss of information (as will be demonstrated by the simulations in Chapter 3). On the other hand, rating scales with too many options (e.g., 1 – 100) likely present significant psychological burden on participants. A final note to mention is that pair-wise rating of n items requires n x (n-1)/2 judgments, which quickly becomes unmanageable once n surpasses 15 or 20.[3]

1.6.4.2 Triplet comparison tasks

Triplet tasks (e.g. Li et al., 2016) involve displaying three stimuli on the screen for participants to make a similarity judgment. From here, they can be subdivided into *triplet matching* task or *triplet discrimination* task (Demiralp et al., 2014). For the triplet matching task, one of the three items is presented as a *query* while the other two are designated as *referents* and the participant is asked to choose which of the referents is more similar to the query item. For triplet discrimination tasks, participants are simply asked to choose the odd-one-out, that is, which of the three items is the least similar to the other two.

---

[3] An additional consideration when using pair-wise judgment tasks is whether to use a similarity or a dissimilarity scale. Although some previous studies have found strong negative correlations between similarity and dissimilarity judgments (Hosman & Künnapas, 1972; Tversky, 1977), implying that similarity = 1/dissimilarity, others have found systematic differences depending on the task and context (Mathy et al., 2013; Medin et al., 1990; Tversky, 1977; Tversky & Gati, 1982).

Some advantages of such tasks are that for any pair of items, their similarity is judged in the context of the third item. Also, only a binary response is required (one of the two referents), which bypasses issues to do with changes in the criteria and the ranges of ratings associated with pair-wise estimations (akin to advantages of forced choice versus yes-no tests in signal detection theory). This typically results in higher within and across participant consistency in judgments than pair-wise similarity ratings (Demiralp et al., 2014; Li et al., 2016). A major disadvantage, however, stems from the time complexity of data acquisition. For example, exhaustively sampling all the trials for the triplet matching task would require n x (n-1) x (n-2)/2 trials.

As reviewed above, various mathematical models have been available to infer low-dimensional *psychological embeddings* of similarity data, such as MDS for pair-wise judgments. In recent years, novel computational models have been developed that use triplet judgments to infer such embeddings and also offer methods to adaptively reduce the number of trials necessary for such inference (e.g. Tamuz et al., 2011). Roads & Mozer (2019) developed the *PsiZ* model which offers these functionalities as a package and which we have used in combination with the triplet matching task in Chapter 2 (also see Roads & Love, 2021 for a discussion of uses of PsiZ).

1.6.4.3 Confusability and identification tasks

Similarity between two stimuli can be assessed by measuring how confusable they are with each other. In a typical *same-different* paradigm, pairs of items are presented (either sequentially or simultaneously) and a participant responds with a “same” or “different” button (e.g. Corter, 1987, experiments 4 and 5; Rothkopf, 1957). Either accuracy or reaction times can be used as proxies for similarity. In identification tasks, a single stimulus would be presented briefly, and a participant would have to identify it using a limited set of available response options (Corter, 1987, experiment 6). Here, similarity between a pair of stimuli is the probability of cross-identification between them.

One limitation of such approaches is that the binary nature of response options limits the precision of similarity judgments. For example, participants can only respond with two options on a same-different task, requiring multiple repetition of trials to get a reasonable estimate of confusability. Additionally, identification tasks involve an implicit choice process at the responding stage of the trial. This makes it difficult to use identification tasks to measure the impact of various experimental manipulations on underlying

psychological similarities, since the final response output is a combination of similarity computation and the choice process (see Corter, 1987 for a discussion).

1.7 A note on statistics in this thesis

Before proceeding to the empirical chapters of the thesis, it is worth outlining the statistical approach that we took in many of the experiments, which employed Bayes Factors (BFs; Dienes, 2016). This was motivated by two considerations: (1) BFs quantify support for the null as well as the alternative hypothesis, and (2) they can be used for stopping criteria in *sequential designs*, which enable more efficient data collection methods (Schönbrodt & Wagenmakers, 2018).

Frequentist traditions do not formally allow quantification of evidence against the null hypothesis, because the p-value provides the probability of obtaining a statistic (given the data) as high or higher than some threshold, *assuming* that the null hypothesis is true. In other words, p-values above some convention (e.g. p>0.05) provide absence of evidence, rather than evidence of absence. With Bayes Factors, however, one can quantify evidence in support or against one hypothesis relative to another in terms of the ratio of likelihoods:

$$BF_{10} = \frac{p(data|H1)}{p(data|H0)} = {}^{1}/_{BF_{01}}$$

where BF_{10} is the Bayes Factor in support of the alternative over the null, and BF_{01} is the evidence in support of the null over the alternative.

Even if one assumes that H1 and H0 are equally likely a priori, the calculation of these two likelihoods requires choices for priors on the parameters that define H1 and H0. Throughout this thesis, we opted to use *objective priors* which depend only the particular statistical procedure used (e.g., t-tests and ANOVAs), i.e. are the same for all tests, regardless of any previous data relevant to H1 or H0 (so-called *subjective priors*; Stone, 2013). For a t-test, we used the *ttestBF* function in the R BayesFactor package (Morey & Rouder, 2021), with the default *JSZ* prior corresponding to the *rscale* parameter of $2/\sqrt{2}$. For ANOVAs, we used the *anovaBF* function from the same package, with default priors corresponding to the *rscaleFixed* parameter at *1/2*.

Unlike in the frequentist tradition, where a relative consensus is reached regarding the threshold for the p-value (i.e. $\alpha = 0.05$), there has been less discussion (i.e. less time for convention to emerge) regarding how large a Bayes Factor needs to be, in order to be considered as "strong" evidence in support of H0 or H1. Typically, journals consider

BF10 > 6 and BF10 < 1/6, or BF10 > 10 and BF10 < 1/10 as "publishable" evidence, and so we have adopted the following convention to interpreting BFs (Jeffreys, 1998; Kass & Raftery, 1995; Quent, 2021):

Table 1.1: Interpreting the strength of Bayes Factor evidence.

BF10	Evidence
> 100	Extreme evidence for H1
30 – 100	Very strong evidence for H1
10 – 30	Strong evidence for H1
6- 10	Moderate evidence for H1
3 – 6	Anecdotal evidence for H1
3 – 1/3	Inconclusive evidence
1/3 – 1/6	Anecdotal evidence for H0
1/6 – 1/10	Moderate evidence for H0
1/10 – 1/30	Strong evidence for H0
1/30 – 1/100	Very strong evidence for H0
< 1/100	Extreme evidence for H0

Finally, within the Bayesian statistical framework, efficient data acquisition methods have been developed that allow termination of data collection in case of early support for one of the hypotheses (Schönbrodt & Wagenmakers, 2018). This can often result in massive savings in time and resources. In Chapters 2, 4 and 5, we used a variation of such an approach called *Bayesian sequential designs with maximal N*, which is described below.

In such designs, data acquisition starts with an initial batch, from which a BF is estimated. If either BF10 or BF01 already exceed a predetermined threshold, then data collection is stopped. Otherwise, another batch of participants of size n (typically a multiple of counterbalancing conditions) is collected, and the (pooled) BFs are checked again. This continues until either one of the BFs exceeds a threshold, or until the maximum number of participants is reached. The latter is normally determined by available resources (e.g.

time or money). Parameters such as the initial number of participants, size of additional batches, thresholds for BF10 and BF01, and the maximum N are predetermined before data collection (and here, pre-registered on OSF for several of our experiment, as cited at the relevant points in the chapters). For brevity, in chapters where I used such a sequential design, I only specify these parameters and the nature of the statistical test used in BF calculation, instead of reiterating the full procedure.

Finally, where we have used this sequential design, we have also conducted simulations for calculating a Bayesian equivalent of "power" for a specific sequential design to support a certain hypothesis. To facilitate easy calculation of such "power" for various setups of Bayesian sequential designs, we created an R codebase which is available on the GitHub: https://github.com/MRC-CBU/cbu_bayesian_sequential_designs

A basic skeleton of such power calculation is described below, with experiment-specific details provided in appropriate sections of subsequent chapters.

To calculate "power" for correctly supporting the alternative hypothesis, 10,000 simulations are performed using the specific sequential design with pre-set parameters, with each simulation representing a hypothetical experiment. For each simulation, an initial group of data points are sampled from a distribution corresponding to the assumed effect size (typically a medium effect size of Cohen's d = 0.5) and BFs are checked. If pre-specified BF thresholds are exceeded, data collection stops, and the outcome is recorded as either supporting the null or the alternative. Otherwise, an additional batch of n data points are drawn from the same underlying distribution and the BFs are recalculated on the combined data. For each simulation, this continues until the BF thresholds are exceeded or until the maximum N is reached. The "power" of this procedure to correctly support H1 when a true effect exists is the percentage of such simulations that resulted in BF10 exceeding the threshold.

To calculate the power for supporting the null hypothesis, another set of 10,000 simulations are performed, this time drawing data from an underlying distribution assuming no effect (i.e. d = 0). The "power" in this case is the percentage of simulations where BF01 exceeded the threshold.

2 SYMMETRY AND THE DISTANCE-DENSITY GEOMETRIC MODEL

2.1 Introduction

As introduced in the previous chapter, classical geometric models of knowledge representation must adhere to axioms of minimality, symmetry, the triangle inequality and segmental additivity. This chapter discusses violations of symmetry, reviews the suggested solutions to these violations, and tests a basic prediction of one of the suggested solutions: the distance-density model of Krumhansl (1978). Briefly, the results did not show support for the distance-density model, questioning whether incorporation of density can account for violations of symmetry, and hence questioning geometric models of knowledge representation.

2.1.1 Violations of symmetry

The symmetry assumption requires that the distance between *a* and *b* be equal to the distance between *b* and *a*: $d(a,b) = d(b,a)$ (Tversky, 1977). As discussed in the introductory chapter, much evidence has amassed that similarity judgments violate this property. When similarity comparison statements take a directional form of "assess the degree to which *a* is similar to *b*", asymmetries arise when objects *a* and *b* swap places. This has been found for perceptual stimuli, such as geometric forms, as well as conceptual ones such as countries. For example, Tversky (1977) showed that North Korea is rated as more similar to Red China than the other way around. In explanation for such asymmetries, Tversky argued for the "focusing hypothesis", that in directional similarity statements, a larger emphasis comes on subject *a* than referent *b*. When combined with his feature-matching contrast model for similarity calculation (Equation 1.2), Tversky showed this accounted for asymmetries, and argued against the validity of geometric models.

2.1.2 The distance-density model and its empirical tests

Instead of dismissing geometric theories, Krumhansl (1978) proposed an augmented model that could account for violations of symmetry:

Equation 1.4: $$\boldsymbol{d'(a,b) = d(a,b) + \alpha \times D(a) + \beta \times D(b)}$$

Here, *d(a,b)* is the metric distance, *D(a)* and *D(b)* are some measures of local density in regions of *a* and *b*, α and β are positive constants, and *d'(a,b)* is the final resulting dissimilarity. Thus, final dissimilarity is a function of some "objective" distance in a geometric space plus densities around the points. If $\alpha > \beta$ due to the focusing hypothesis, then $d'(a,b) > d'(b,a)$ whenever $D(a) > D(b)$. A more basic prediction of the model is that density stretches the psychological space between points.

Does the distance-density model stand up to empirical scrutiny? In support, Krumhansl (1978) presented re-analysis and re-interpretation of previous experimental data. For example, Rothkopf (1957) used a same-different task on Morse code signals and found large asymmetries as a function of the ordering of signals. Re-analysing the data, Krumhansl pointed out that for those pairs with large asymmetries, the similarities tended to be smaller when the first stimulus was the one with more neighbours in a nonmetric MDS solution of the data (analysed separately by Shepard, 1963), i.e. was in a denser region. For Tversky's data on countries and figures, prominent stimuli (like "Red China") tend to have more features, meaning a larger number of neighbours sharing those features, and thus higher density. Finally, for the asymmetries documented by Rosch (1975) with prototypical and non-prototypical stimuli, Krumhansl argued that when scaling solutions are applied to stimuli, prototypes tend to be placed in denser regions.

Krumhansl recognised the need for directed and targeted experiments that test effects of density on similarity, instead of relying on just re-analysing data. The first intentional tests of the distance-density model came from Corter (1987), who failed to find any effect of density on similarity. Corter used a between-participant manipulation of density by presenting more neighbours for certain target stimuli, expecting this increased density to produce a reduction in overall similarity of these targets to all other stimuli. For all of Corter's experiments, the manipulation of density involved the participants simply flipping through a booklet containing images of every exemplar and indicating once they were familiarized with all items. For the first 5 of the experiments, this was followed by non-directional pair-wise similarity judgments or same-different judgments to assess any stretching of psychological space around denser regions. Across the first three

experiments with pair-wise judgments of ellipses, faces, or letter-like stimuli, no such changes in similarity ratings were found. In two other experiments using the same-different task with letters or letter-like stimuli, Corter again found no effects of density. In the 6th experiment, with an identification task used on letter stimuli, Corter did find effects on density but attributed this to the choice process during the selection of a response, not to the underlying similarity computation on the representations. Thus, even the simplest prediction of the distance-density model – that density around a stimulus should impact its overall similarity with other items – did not pan out (although see Krumhansl's 1988 comment on these experiments and Corter's 1988 reply to the comment).

2.1.3 Types of density manipulations

Before evaluating Corter's claims, it is worth considering two distinct ways of experimentally increasing density in a particular stimulus region. Each experience with an exemplar creates a psychological imprint at the corresponding coordinate in psychological space. However, this imprint is not precise, but spreads to neighbouring locations due to inaccuracy in encoding into (or possibly forgetting from) memory. This spread is often interpreted as the probability of generalising a stimulus response to neighbouring stimuli due to similarity between the stimuli (e.g. Nosofsky, 1986; Shepard, 1957, 1987). This similarity can be modelled as a Laplacian kernel with an inverse exponential on both sides described by the formula:

Equation 2.1: $$s = e^{-\beta d}$$

where s is the similarity, d is the (unsigned) distance in physical space, and β is a parameter governing the gradient. To increase density around that exemplar, one can either repeat the same exemplar or present neighbouring stimuli. Assuming additive density, both methods should result in larger density around the exemplar in the psychological space (Figure 2.1).

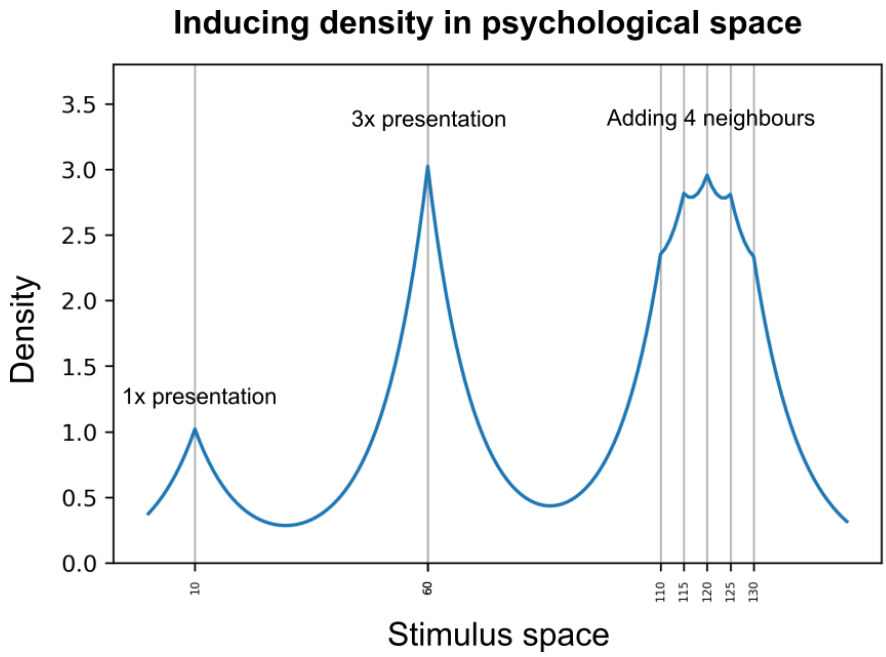


Figure 2.1: Two ways of inducing density in psychological space.
A single experience with a stimulus exemplar is thought to produce a psychological imprint, increasing the psychological density at the stimulus coordinate. Repetitions of the stimulus is thought to add to total density at and around the stimulus coordinate. Alternatively, neighbourhood density can be increased by keeping individual stimulus repetition frequency the same but presenting neighbouring stimuli, which additively drive up density.

While Corter (1987) employed the second method of adding neighbours, Polk and colleagues (2002) manipulated density by re-exposure to the same exemplars multiple times, and found that asymmetries in similarity judgments *could* be directly induced. In a within-participant manipulation, participants rated directional similarity statements (e.g. "how similar is *a* to *b*") between colour patch stimuli before and after an exposure task. During the exposure task, they performed a size-judgment task where colour was irrelevant but where certain colour patches were presented ten times more than others. The authors showed a significant increase in asymmetric similarity judgments post- compared to pre-exposure. The authors argued that, even when number of features are held constant (colour), increasing salience of a stimulus through frequency will lead to asymmetric judgments. It is worth noting that, as the colour feature was fully orthogonal to the size-judgment task, the exposure task did not involve any perceptual training, which is sometimes hypothesized to be a mechanism underlying changes in similarity due to experience (Corter, 1987). Finally, Polk et al. (2002) found an overall global increase in similarity ratings (averaged over both directions) post- versus pre-exposure task, although

they did not report if this increase was larger for those judgments involving manipulated colour patches versus those without such patches, as would have been predicted by the distance-density model.

Although Polk and colleagues did not consider this, if density is modelled as in Figure 2.1, stimulus repetitions could result in increases in neighbourhood densities. Thus, their results could be taken as supporting evidence for the distance-density model. However, such saliency-induced asymmetries can be equally accounted for by the feature-based contrast model proposed by Tversky (1977), where similarity between items *a* and *b* is a function of their corresponding feature sets *A* and *B*:

Equation 1.2: $\boldsymbol{S(A,B) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)}$

Here, the scale parameter *f* captures the salience or prominence of features. Multiple repetitions of item *a* would increase the salience of its features *f(A)* and thus of *f(A – B)*. In directional similarity judgment tasks when $\alpha > \beta$, such an increase in *f(A – B)* will disproportionately influence *S(A,B)* when the salient stimulus is the subject of the comparison judgment, as opposed to *S(B,A)* when it is the referent. This, in turn, will result in *S(A,B) < S(B,A)*.

Importantly, the feature-matching model has no way to account for influences from neighbouring stimuli, which makes the manipulation of density through presentation of neighbouring stimuli as the proper way to tease it apart from the distance-density model. This is the approach we took in this Chapter.

2.1.4 The current experiment

Although Corter (1987) varied neighbourhood densities of his stimuli, the changes might not have been strong enough to elicit a detectable difference in density. In this chapter, we set out to demonstrate effects of density with stimuli comparable to those used by Corter (i.e. varying in physical shape) but involving a stronger manipulation of neighbourhood density. We created a novel one-dimensional artificial stimulus space (see Figure 2.2). We then ran an initial Norming study using the *triplet matching task* which involved presentation of 3 exemplars, one of which was a *query* item while the other two were *referents*, and where the participants indicated which of the referents was more similar to the query item. This allowed us to employ the PsiZ model for inferring psychological embeddings, which let us confirm that exemplars sampled from our space were roughly equally far apart in psychological space. This would mean that the

relationship between the *generative* space (i.e. the physical characteristics) and the *psychological* space (i.e. internal representations) was roughly linear.

In the main experiment, we used a pre-post design with an exposure task in between, similar to Polk et al. (2002). Unlike Polk et al. (2002) or Corter (1987), we used the triplet matching task to assess the pre- and post-exposure similarities, primarily because this would allow us to fit the PsiZ model in subsequent iterations of the paradigm, but also because triplet tasks are less noisy than pair-wise judgments (Demiralp et al., 2014; Li et al., 2016). For the exposure task, we manipulated the neighbourhood density of exemplars but, unlike Corter (1987), we added a substantially larger number of neighbours using a within-participant design, which should increase statistical power. The task involved a variation of a same-different judgment paradigm, with stimulus exemplars presented on the screen one after another, and the participants asked to compare the on-screen exemplar with the preceding one. Therefore, instead of a passive viewing task as used by Corter (1987), our task involved active judgment, ensuring participant engagement and increasing the chances of successfully modifying the psychological density. Furthermore, as discussed in Chapter 1, confusion probabilities can be used to assess similarities between exemplars. Thus, unlike the exposure task of Polk et al. (2002), our task provided another way of assessing impact of density on similarity, via confusion probabilities between stimuli.

Finally, when employing any stimulus space in an experiment, one inadvertently creates sharp boundaries at the edges of the distribution. Krumhansl argued that boundary stimuli are located in less dense regions, which she speculated could explain some results in the literature, such as the finding that self-similarities of boundary exemplars tend to be larger than those of non-boundary stimuli (Krumhansl, 1978). Differences in density at boundaries also offer a chance to test another prediction of the distance-density model: a pair of stimuli in which one is a boundary exemplar should look more similar than another pair in which both items are non-boundary stimuli. Therefore, we looked for boundary effects in our 1D space by comparing the choice probabilities associated with boundary vs non-boundary referents in our triplet matching task.

2.2 Norming study

To ensure a roughly linear relationship between the generative and psychological spaces of our 1D stimuli, we conducted a norming study where a separate group of participants did a triplet matching task on multiple exemplars sampled from the space. These

judgments were then passed to the PsiZ model to estimate psychological embeddings of these exemplars, i.e. distances between the internal representations of the stimuli. The model additionally allowed us to estimate the β parameter underlying the assumed exponential similarity function, which was then used to predict the impact of our density manipulation in the main experiment, as explained below.

2.2.1 Methods

2.2.1.1 Participants

A total of 13 healthy young adult participants were recruited (10 females) from the prolific.co platform, aged 19-44 (M = 31.38, SD = 7.14), and paid £6/hour for their time, according to the Cambridge Psychology Research Ethics Committee protocol PRE.2020.018. Of these, 11 (8 females) aged 19-46 (M = 31.1, SD = 7.73) passed the final quality and performance checks (see the *Quality checks* section below) to be included in the data analysis.

2.2.1.2 Stimuli

We used the Blender software (Blender Online Community, 2018, version 2.82a) to design an artificial stimulus space with exemplars consisting of a cone and a triangular base. Exemplars differed by the shape of their base, varying from circle-like convex shape to concave ones (see Figure 2.2). For the norming study, 11 evenly spaced stimuli were chosen between exemplars 20 and 120. The dimensions of stimuli as displayed during the experiment varied between 186x282 pixels for the most convex stimulus and 240x330 pixels for the most concave one. Note that due to online nature of the experiment, the visual angle subtended by stimuli would have varied for each participant depending on the screen resolution and distance to the screen.

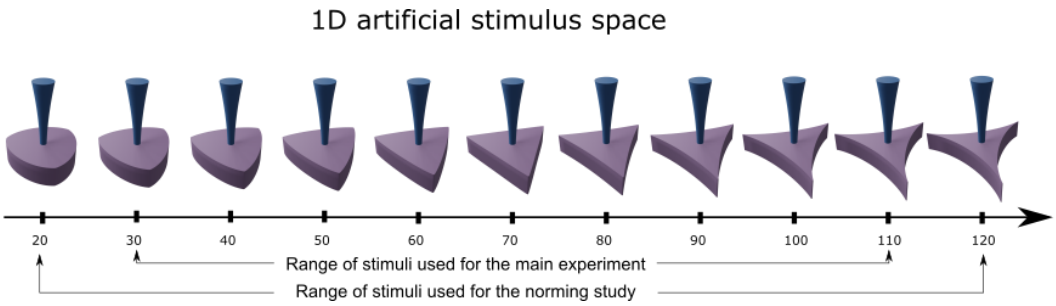


Figure 2.2: 1D artificial stimulus space.
Stimuli varied by curvature of the sides of their base component. 11 stimuli between coordinates 20 – 120 were used for the Norming study, subsequently reducing this range to 30 – 110 for the main Experiment.

2.2.1.3 The triplet matching task

The triplet matching task involved making a relative similarity judgment between three exemplars displayed on a screen. One of the exemplars was the *query* while the other two were *referents*. Participants had to respond by choosing which of the referents looked more similar to the query. See Figure 2.3 for the task design and an example trial.

The participants started with a short practice block, consisting of 10 trials using 10 exemplars different from those used in the main experimental blocks. On each trial, the participants used their keyboards to press either "q" or "p" to indicate whether the left or the right referent was more similar to the query, respectively. The triplets stayed on the screen for a minimum of 2 seconds to discourage rapid, mindless responding and a maximum of 5 seconds if no response was given, after which a blank inter-trial-interval (ITI) of 500ms ensued, followed by the beginning of the next trial. Due to the subjective nature of similarity, trial-specific feedback was not given. Instead, a block-specific average accuracy was presented at the end of practice trials, calculated by taking accuracy on those trials that had a "correct" response based on distances in generative space.

During the main blocks, all 495 possible triplets consisting of all combinations of the 11 exemplars were presented to the participants, split over 6 blocks (83 trials per first 5 blocks, 80 trials for the 6^{th}). The trial structure was identical to the practice trials, with block-specific average accuracy feedback given after each block. The participants were informed of a potential bonus payment depending on the quality of their responses (bonus was up to £1.5, proportional to total accuracy across all trials).

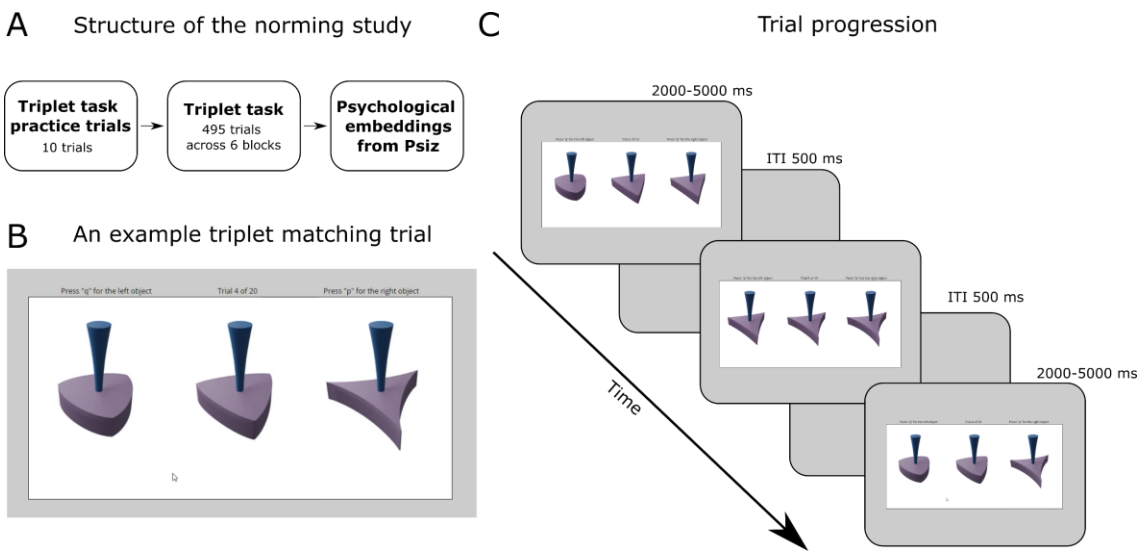


Figure 2.3: The norming study.

(A) Overall structure of the study. Participants started with practice trials and then completed 6 blocks of triplet matching trials. Data was subsequently fed to PsiZ for estimating psychological embeddings. (B) An example triple matching trial. (C) Trial progression during the norming study. Every trial was followed by an ITI of a blank screen displayed for 500ms.

2.2.1.4 Quality checks

Participants were excluded if they reported any technical difficulties or not having understood the task instructions. A total of two participants were excluded due to technical issues (images not loading and PC restarting due to updates).

2.2.1.5 Inferring psychological embeddings with the PsiZ model

The PsiZ model allows inference about psychological embeddings, which consist of multi-dimensional feature representations of stimuli and a corresponding similarity function that describes the degree to which the response associated with one stimulus transfers to another (Nosofsky, 1986; Shepard, 1987; Tenenbaum, 1999). For full details about the underlying generative model and the inference procedure see Roads and Mozer (2019).

We chose the *Laplacian kernel* as our similarity function, whereby similarity between two points in a two dimensional space (with coordinates a, b) is an exponentially decaying function of the Euclidean distance between them, i.e, Equation 2.1 with a Euclidean definition of distance: $s(a,b) = \exp(-\beta * \sqrt[2]{a^2 - b^2})$. Although PsiZ allows joint estimation of the embedding locations and of β, the two parameters trade off against each other and might result in unnecessarily long convergence time. Therefore, we opted

for a two-step approach: (1) fix β to 10 and infer embedding coordinates to test for linear relationship with the generative space, and (2) if the relationship is linear, estimate the β parameter by fixing the embedding coordinates.

2.2.2 Results

2.2.2.1 Psychological and generative spaces are linearly related

Figure 2.4 shows the relationship of the distances between our exemplars in the generative space and their distances in the psychological space, using group-averaged data and β = 10. The results show a strikingly linear relationship, indicating psychological uniformity along our stimulus dimension and validating our stimulus set for subsequent experimentation. However, the first stimulus (exemplar 20) showed a large error bar, so we removed this exemplar from subsequent experiments.

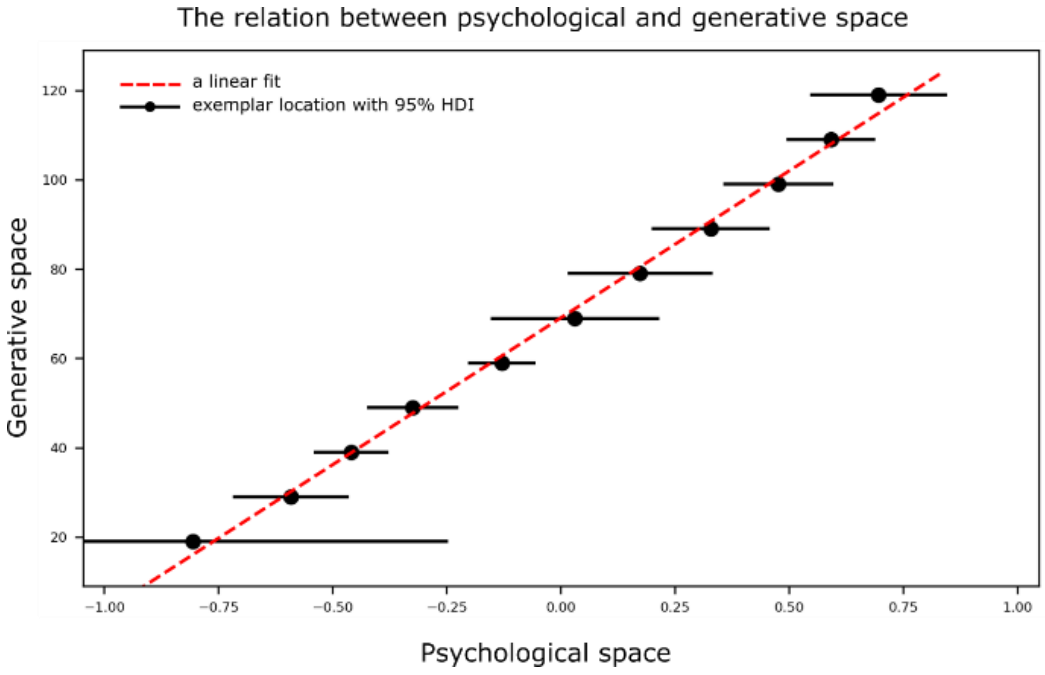


Figure 2.4: Psychological embeddings of the 1D stimuli from the Norming study. *x* axis corresponds to locations in psychological space as estimated by PsiZ. *y* axis corresponds to locations of the 11 stimuli in the generative space. Each dot is an exemplar stimulus. Error bars indicate 95% highest density interval (HDI), denoting the variance in the posterior estimate of the location of the points in psychological space. The red line is a linear fit through the data.

2.2.2.2 The β parameter for the similarity function

Setting the embedding coordinates equal to those in the generative space, we estimated the β parameter to be 0.14. This was used in our subsequent experiment to predict density changes as a result of our manipulations.

2.3 Experiment

We tested the most basic prediction of the distance-density model: whether increased exposure to exemplars from one part of our 1D stimulus space (i.e. increased "density" in that part) would result in changes in perceived similarity between exemplars from that part of the space. This should influence behaviour on the triplet task: for a given triplet where one referent is in the *low-density* region of the space, the other is in the *high-density* region, and the query item is between the referents, the post-exposure psychological distance between the query and the high-density referent should increase, leading to a lower probability of choosing that referent (Figure 2.5).

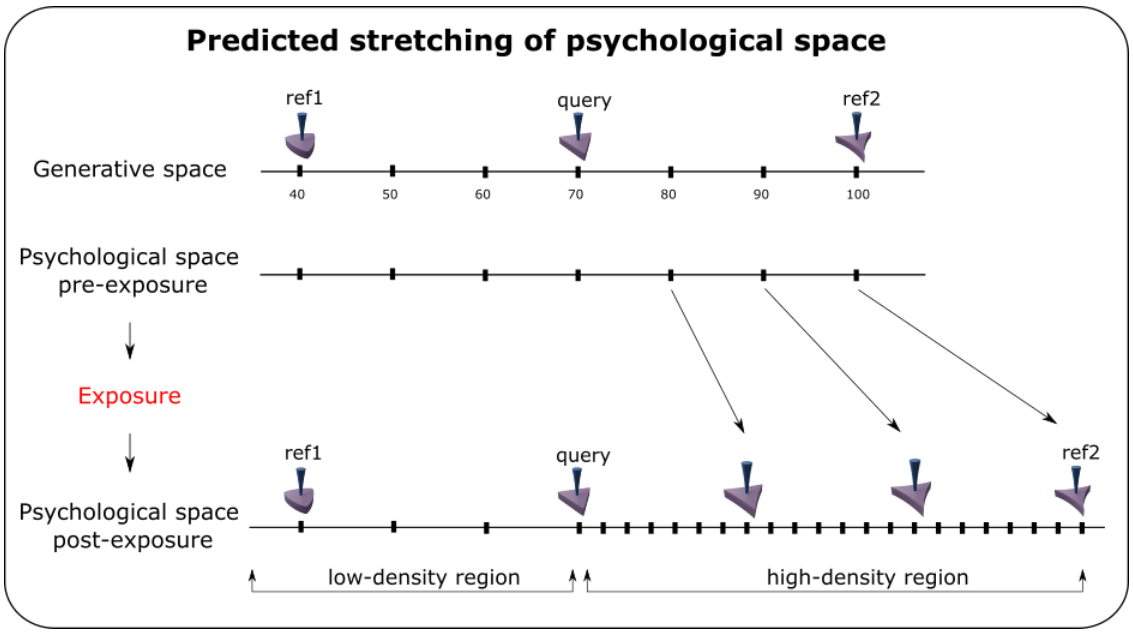


Figure 2.5: Predicted stretching of psychological space due to density.
The top panel depicts physical generative space, where for a symmetric triplet the referents are equally distant from the query item. This translates to equal distances in psychological space as well. During the exposure phase, more exemplars are displayed on one part of the space compared to another, creating high-density and low-density regions. This stretches the psychological space in the high-density region, leading to referent 2 being further away from the query than referent 1.

2.3.1 Methods

2.3.1.1 Participants

A total of 141 healthy young adult participants were recruited (82 females) from the prolific.co platform, aged 19-46 (M = 32.1, SD = 6.87), and paid £6/hour for their time, according to the Cambridge Psychology Research Ethics Committee protocol PRE.2020.018. Of these, 110 (61 females, 78% of those recruited) aged 19-46 (M = 31.7, SD = 6.9) passed the final quality and performance checks (see the *Quality and performance checks* section below).

2.3.1.2 Stimuli

We used exemplars between points 30 and 110 from our 1D stimulus set as described above for the Norming study (Figure 2.2). The stimulus at coordinate 70 corresponded to an exemplar with a base with straight edges. Thus, the region of our stimulus space below 70 was denoted as the *convex* region whereas above 70 was denoted as the *concave* region according to the shape of the base segments.

2.3.1.3 Modelling changes in psychological density

As discussed in the Introduction, we modelled changes in psychological density as a result of exposure to exemplars with Equation 2.1 (and β=0.14), and assumed additivity of density such that exposure to the same or nearby exemplars would summate linearly. For simplicity, no memory loss was assumed across time, such that the initial imprint remained constant throughout our short experiment.

2.3.1.4 Task design and procedure

The participants did two triplet matching tasks, separated by a same-different task that served to moderate the density in parts of the stimulus space (Figure 2.6).

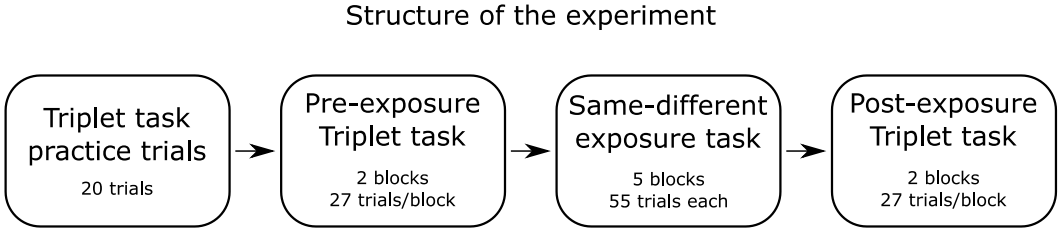


Figure 2.6: The task progression for the main experiment.

Participants started with practice trials to get accustomed to the trial structure. After this, the participants performed pre-exposure triplet matching task, which was followed by the same-different exposure task where more exemplars were

presented from one side of the space. Finally, a post-exposure triplet task was administered. Breaks occurred between tasks and blocks.

2.3.1.4.1 The triplet matching task

The similarity task and trial structure were same as described above for the Norming study. Participants started with a short practice block of 20 trials, with block-specific feedback of average accuracy given at the end. Since these trials could already induce some density in specific parts of the psychological space, we chose these triplet trials such as to ensure a relatively uniform density distribution that was symmetric around the mid-point 70. See Figure 2.7-A for a plot of the density distribution post-practice trials, assuming $\beta = 0.14$.

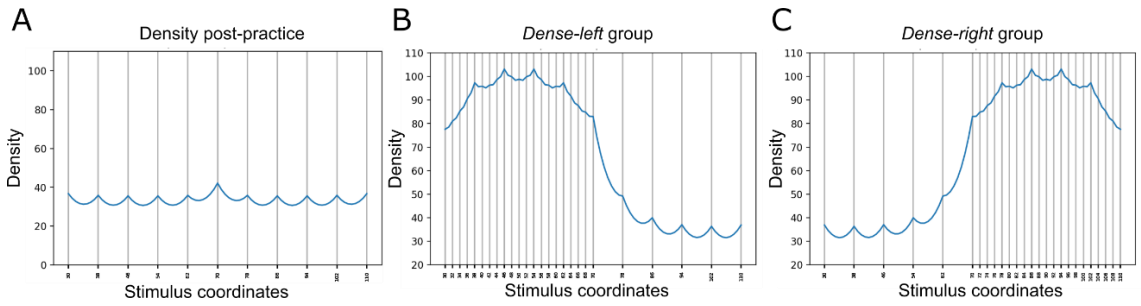


Figure 2.7: Predicted psychological densities at different stages of the experiment. (A) Relatively uniform psychological density after practice trials. (B) Psychological density for the Dense-Left group of participants that were exposed to more exemplars on the left side of the 1D stimulus space compared to the right side. (C) Psychological density for the Dense-Right group of participants. Vertical grid lines indicate coordinates of the exemplars that were presented.

For the main blocks, instead of generating all combinations of specific exemplars, we used triplets consisting of exemplars that were evenly sampled from the low-density, middle, and high-density regions of our 1D space, as described below. A total of 27 triplets were chosen. For each triplet, the coordinate of the query item was between the coordinates of the two referents. Both the pre and post triplet tasks were split into two blocks, such that each of the 27 triplet trials were repeated twice, counterbalancing which of the referents was presented on the left or the right side of the screen. Block-specific feedback was given at the end of each block, consisting of average accuracy on those trials with valid "correct" response in generative physical space.

2.3.1.4.2 The same-different "exposure" task

To induce density in parts of the psychological space, we needed to systematically expose our participants to exemplars. Instead of doing more triplet trials, where one exemplar (the query) might carry more psychological weight, we opted for a more neutral 1-back same-different task (Figure 2.8). On each trial, an exemplar was shown on the screen and the participants pressed "q" if they thought the exemplar was different from the previous one, and "p" if they thought it was the same. Each trial lasted until a response was given, or for maximum of 3 seconds, after which the next trial began. Regardless of the response, the stimulus stayed on the screen for at least 2 seconds to help ensure it was properly processed. The ITI was 500ms. To avoid any after-effects, and reduce trial-to-trial perceptual influences, a grey square was used as a mask in between exemplar presentations, and each exemplar appeared at a randomly jittered location on the screen. The 260 trials were split up into five blocks of 52 trials, with breaks in between. Feedback was given only at the end of each block, summarizing the averaged performance for that block.

Half of the participants (*Dense-Right* group) saw more exemplars from the right-hand side of the space, whereas the other half (*Dense-Left* group) saw more exemplars from the left-hand side (see Figure 2.7 panels B and C for exemplars chosen). Six exemplars (including exemplar 70) were taken from the low-density part of the space and 20 from the high-density part. Each exemplar was repeated 10 times, resulting in a total of 260 trials. The final sequence was pseudo-randomized ensuring at least 20% of the trials contained the same exemplar as the one before, to keep the participants engaged in the same-different judgments. Figure 2.7 panels A and B shows the modelled density in the psychological space resulting from the practice trials and exposure trials for each counterbalancing group, assuming the exponential similarity function with parameter $\beta = 0.14$ as estimated during the Norming study.

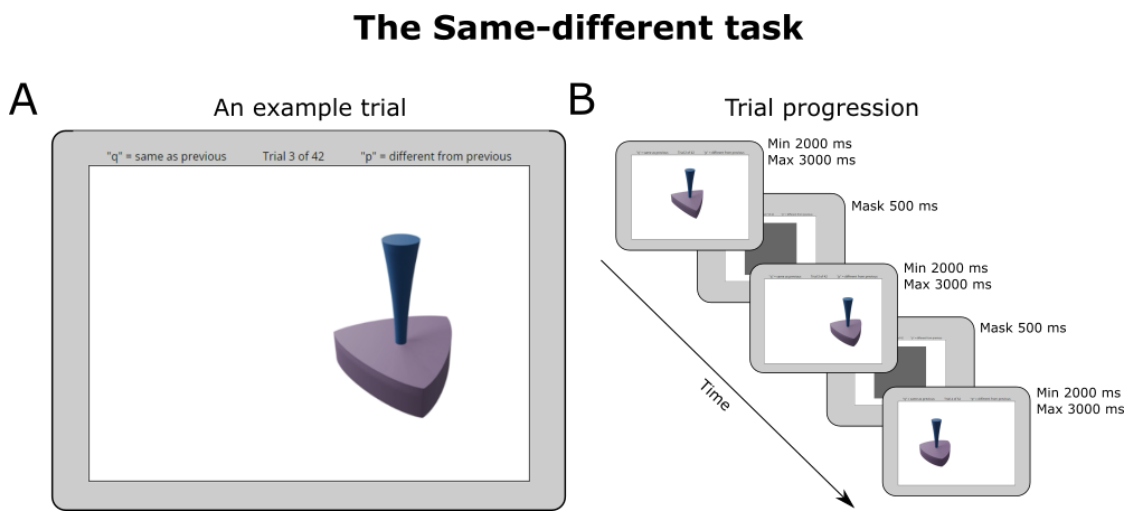


Figure 2.8: Inducing psychological density with the same-different "exposure" task. (A) An example snapshot of a same-different judgment trial. Participants had to compare the current stimulus with the previous one. Regardless of the response, stimulus remained on the screen for at least 2 seconds to ensure adequate exposure. (B) Trial progression. Every trial was followed by a 500ms ITI during which a grey mask was presented. On every trial, the location of the stimulus on the screen was varied randomly.

2.3.1.5 Quality and performance checks

A participant was excluded from analysis if any of the following occurred:

- For the triplet matching task, the same button was pressed too many times in a row, indicating low engagement and attention to the task. The cut-offs for this procedure are explained below.
- Within any block of the triplet matching task, the combined number of missed trials and trials with RT<300ms exceeded 20%.
- For the "easiest" trials of the triplet matching task, i.e. those where the difference between distance(query,ref1) and distance(query,ref2) are maximal, more than 25% of the responses were wrong. The correctness of the trial was defined relative to the generative space.
- Any of the breaks lasted longer than 10 minutes.
- Debriefing surveys indicated presence of any technical difficulties during the task.
- The participant reported to have misunderstood the instructions.

2.3.1.5.1 Determining the cut-off for sequential button presses during the triplet matching task

A long sequence of same button responses could be indicative of paying low attention to the task. To determine the threshold for excluding participants on this criterion, we developed the following procedure.

An "ideal observer" data was simulated for every triplet matching trial. For *asymmetric* trials where one referent was closer to the query in generative space than the other referent, the "correct" response corresponded to the closest referent. For *symmetric* trials where each referent was equally distant, response was chosen randomly. The response data were then shuffled 10,000 times. For each of the 10,000 permutations, we counted how many times a sequential button press (either "p" or "q") of length n occurred, i.e. obtaining a probability distribution of expected number of repeats of length n. Having obtained such permuted distributions for the "ideal observer", each participants' data was also counted for number of repeats of length n and excluded if this number fell in the top 5^{th} percentile of the permuted distribution. This procedure was repeated to check for various lengths of sequential button presses, with n between 4 and 10.

2.3.1.6 Data analysis:

2.3.1.6.1 Triplet locations

For the triplet matching task, the 27 triplets were evenly sampled from three regions of our 1D space: the *convex* region, the *middle* region, and the *concave* region. See Figure 2.9 for examples. The triplets in the convex or concave regions contained a query item with a convex or concave base, respectively. The triplets in the middle region consisted of a query item with a straight base edge (exemplar 70), and one referent in the convex region and the other in the concave region.

Given the two counterbalancing groups that were exposed to exemplars from different halves of our 1D space, the triplets in the convex region corresponded to being either in *low-density* (for the Dense-Right group, Figure 2.7-C) or *high-density* (for the Dense-Left group, Figure 2.7-B) regions, and vice-versa for the concave region. The middle triplets always had one referent in the low-density and one in the high-density regions.

2.3.1.6.2 Triplet templates

Triplets from different regions could belong to the same *template*, i.e. have the same distances between the query and references. This would make them equally difficult for

the similarity judgment, regardless of what part of the 1D space they are in. A total of 9 such templates were used, with 3 triplets per template, resulting in 27 triplets (Figure 2.9).

2.3.1.6.3 Triplet easiness

Some of the triplets were *symmetric*, meaning that the two referents were equally distant from the query making these triplets the "hardest" for similarity judgments as no "correct" answer existed relative to the generative space. Other triplets were *asymmetric*, with one referent being closer to the query than the other. Within such asymmetric triplets, we further categorized them by how much closer in generative space one referent was to the query compared to the other referent. Overall, we ended up with three levels of easiness for our triplets: *easy-0* (corresponding to the symmetric ones), *easy-8*, and *easy-16*. See Figure 2.9 for examples.

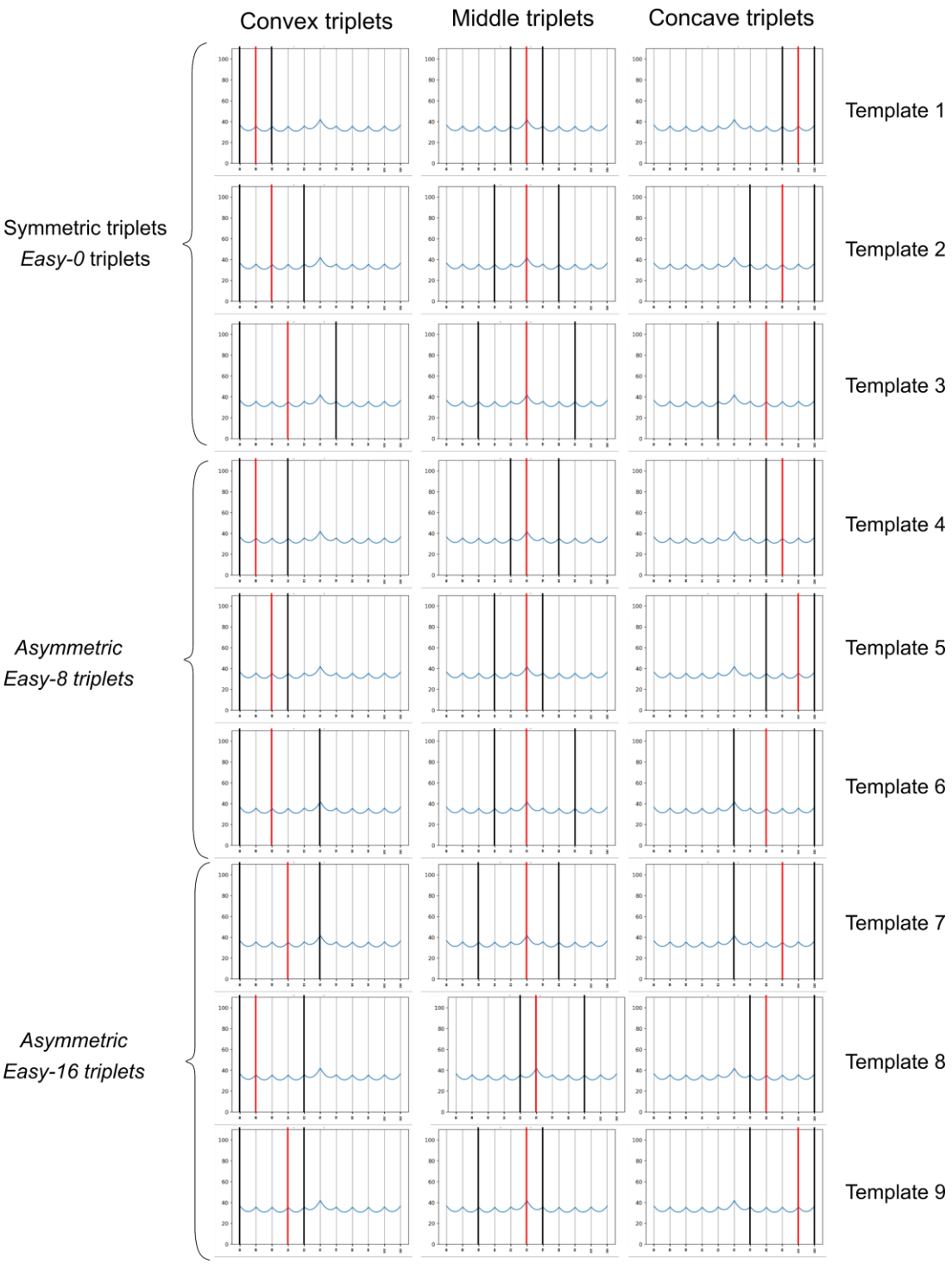


Figure 2.9: The triplets types used for the main experiment.

Red vertical lines denote locations of query items, while black lines indicate coordinates of referents. Blue line is the estimated density after the practice trials. A total of 27 triplets were chosen, belonging to 9 different templates depending on their "shape", i.e. distances between their query and reference items. Each triplet was located in convex, middle, or concave parts of the space based on the position of its query item. Triplets could be symmetric (top 3 rows) or asymmetric (bottom 6 rows), and belong to one of 3 difficulty levels depending on how much closer one

referent was to the query compared to the other referent: Easy-0 (hardest), Easy-8 (medium difficulty) and Easy-16 (easiest).

2.3.1.6.4 Probability of choosing the referent towards the low-density region

Our main confirmatory analysis focused on the *middle symmetric* triplets (Figure 2.9, top 3 of the middle column), where we expected that density would make the high-density referent look further away than before (Figure 2.5), resulting in a higher probability of choosing the low-density referent. We calculated this probability for every triplet pre- and post-exposure task, designating it with a variable *p(chose-low-density)*, expecting a positive change in this variable post compared to pre-exposure.

2.3.1.6.5 Reaction time differences

As one of the exploratory analyses, we looked at reaction times (RT) during the triplet task. The speed of making a choice on a given triplet trial might reflect the easiness of comparing the referents to the query item, which in turn could be driven by the distances between the query and the referents in the psychological space. Thus, RTs might be more sensitive than calculation of choice probabilities for capturing any density-related stretching of the psychological space. We calculated average RTs for each triplet pre- as well as post-exposure, and compared the post-pre difference scores within triplet templates, i.e. compared triplets from different parts of our 1D region.

2.3.1.6.6 Bayesian analysis using sequential design with maximal N

As discussed in the introductory chapter section 1.7, we used a Bayesian sequential design with maximal N procedure to assess evidence in favour of the null or the alternative hypothesis. H1 stated that for the middle symmetric triplets, our main dependent variable of post-pre difference in p(chose-low-density) would be positive, as assessed using a one-sided Bayesian t-test. H0 stated that this difference would not be positive. BF10 and BF01 thresholds were set to 6, initial group size was 20 participants, batch size was 10, and maximum N was 110 giving sufficient "power" to support either H1 or H0 (see below for the power calculation).

We used Spyder (Raybaut, 2009) with Python 3.8 (Guido & Drake, 2009) and RStudio (http://www.rstudio.com/) with R statistical software (R Core Team, 2022) for data preprocessing and analysis.

2.3.1.6.7 Power calculation:

As expanded upon in the introductory chapter section 1.7, we used simulations to calculate "power" of our specific Bayesian sequential design procedure for supporting either H0 or H1. With max N = 110, assuming existence of a large effect (Cohen's d = 0.84), 100% of the simulations resulted in correctly supporting H1 ($BF_{10} > 10$). This effect size was chosen based on a previous pilot data of 20 participants showing a large post-pre difference in p(chose-low-density). In case of an absence of an effect (d = 0), 81.9% of the simulations resulted in correctly supporting H0 ($BF_{01} > 6$), while 1.5% incorrectly supported H1 ($BF_{10} > 10$), and 16.56% remained undecided.

2.3.2 Results: the triplet task

All the analysis and data for the main experiment are available on OSF.

We performed our main confirmatory analysis on the middle symmetric triplets, looking at the post-pre differences in p(chose-low-density). If density stretches the psychological space, the referent in the high-density region should look further away post-exposure for such middle symmetric triplets, leading to a higher probability of choosing the referent in the low-density region.

We ran several exploratory analyses to further examine density effects. First, we looked at post-pre changes in reaction times (RTs) as measures of changes in easiness of making similarity judgments. For any given triplet, a change in post-pre RTs could reflect simple training effects as opposed to density effects. Therefore, to control for generic training effects, we performed *within-template* analysis of post-pre RT difference scores, comparing the RT changes for the middle triplets to those in the convex and concave parts of the space. Second, we also looked at any boundary effects on choice probabilities, whereby the exemplars at the leftmost or rightmost edges of our 1D distribution might have been chosen more often over equally distant referents towards the centre of the distribution. As explained in the introduction, this was motivated by Krumhansl's (1987) proposal that such boundary effects might be explained by the natural lack of neighbouring stimuli (i.e. density) around edges of stimulus spaces.

2.3.2.1 The middle symmetric triplets show no effects of density

Figure 2.10-A shows the post-pre probability of choosing the referent towards the low-density region, in symmetric triplets predicted to be most affected by the manipulation. After acquiring the maximum n of 110 successful participants, Bayes factor for the null

hypothesis of no effect was BF_{01} = 2.41 indicating inconclusive evidence (Cohen's d effect size = 0.12). Figure 2.10-B shows the same effect broken down by the two counterbalancing groups, with BF_{01} of 5.97 and 1.18 for the Dense-Left and the Dense-Right groups, respectively. Thus, contrary to our pilot study where the same comparison yielded a large effect size of d = 0.84, we did not find evidence to support the hypothesis that density changes perceived similarity.

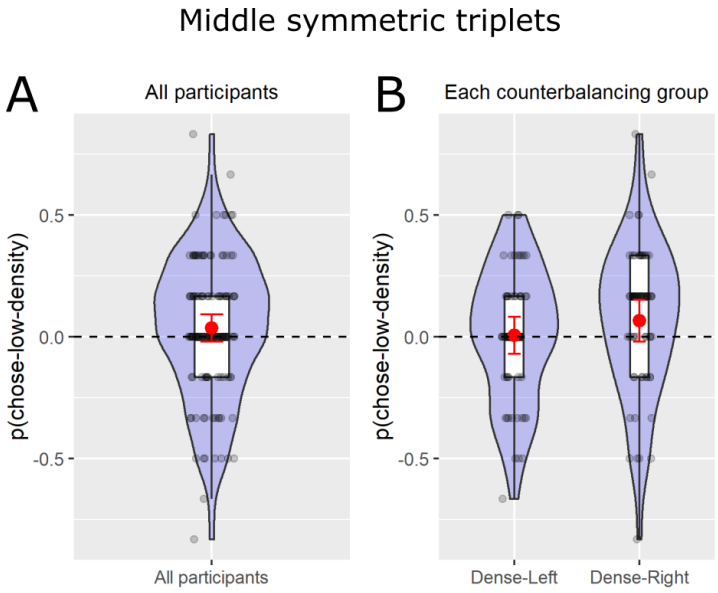


Figure 2.10: Post-pre probability of choosing the low-density referent for middle symmetric triplets.

(A) p(chose-low-density) for all participants. (B) No effect for either of the two counterbalancing groups. Each dot is a participant. Red dots signify group-level means. Error bars are 95% CIs.

2.3.2.2 RT analysis on symmetric triplets reveals overall post-pre training effects but no effect of density

For each participant and each triplet, we calculated RT averaged across the two triplet presentations, separately for the pre- and post-exposure task. Post-pre differences showed consistent speeding-up for both counterbalancing groups (Figure 2.11 A and B), indicative of generic training effects.

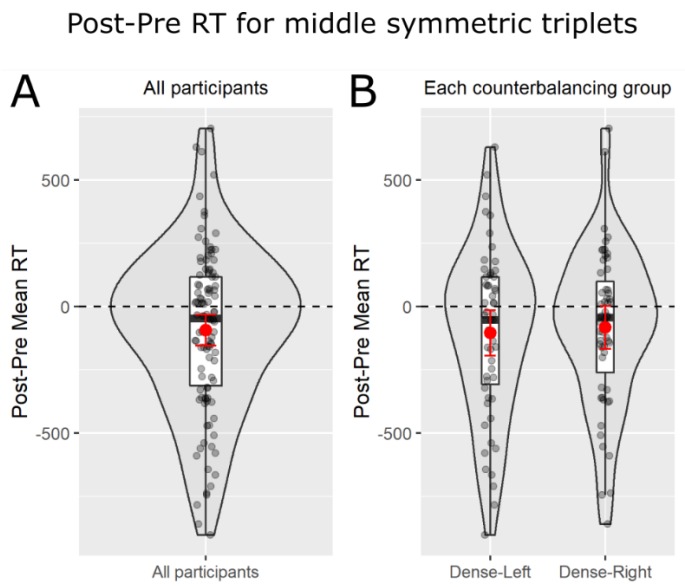


Figure 2.11: Generic training effects shown by RT speed-up.
(A) Post-pre average RT difference for middle symmetric triplets. showed a speed-up effect, BF10 = 8.36. (B) The effect was pronounced for both counterbalancing groups. Each dot is a participant. Red dots signify group-level means. Error bars are 95% CIs.

Generic training effects would impact all triplets from all parts of the 1D space. However, if density affects similarity, middle symmetric triplets should get an additional RT boost as one referent would be perceived as further away. To test this, we ran a within-template comparison of post-pre RT differences for the middle symmetric triplets versus average of high- and low-density region symmetric triplets. Again, we did not find a difference, with BF_{01}=6.34 indicating moderate evidence in support of the null (Figure 2.12).

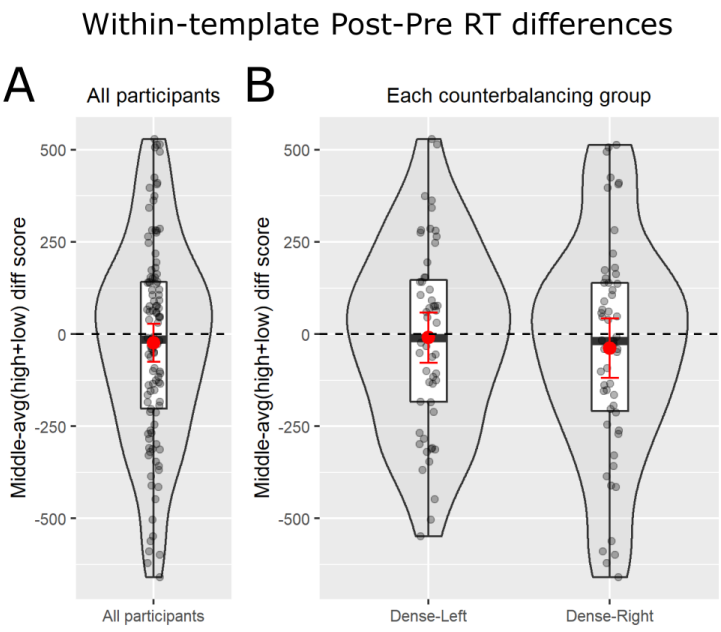


Figure 2.12: Within-template analysis of post-pre RT differences for middle symmetric triplets.

(A) There was no difference for the RT speed-up between the middle triplets and the average of triplets in the high- and low-density regions. (B) Neither of the two counterbalancing groups showed a within-template difference for RT speed-up. Each dot is a participant. Red dots signify group-level means. Error bars are 95% CIs.

2.3.2.3 RT post-pre difference shows a speed-up for asymmetric triplets with the correct referent towards the high-density region.

It is possible that our manipulation created a linear density gradient across the stimulus space, instead of distinct low-density, middle and high-density regions as in Figure 2.7 panel B or C. In that case, the relative stretching between the query and the two referents would be comparable for triplets in all three regions. Due to this, subtracting variables between triplets from different regions would cancel out any effects of density.

Asymmetric triplets offer a solution to this problem. For those asymmetric triplets that have the closest (i.e. "correct" in generative space) referent in the direction of higher density, the post-exposure change in density would stretch the psychological space between the query and the "correct" referent, making the trial harder. This would result in slowing of RTs post- compared to pre-exposure. On the other hand, for the asymmetric triplets with the "correct" referent towards the low-density side (i.e. the "incorrect" referent towards the high-density side), post-exposure RTs should be faster compared to pre-exposure.

Thus, we examined the post-pre RT differences for the above two groups of asymmetric triplets: (i) ones with the closer referent towards the high-density side and (ii) those with the closer referent towards the low-density side. We looked separately at easy-8 and easy-16 asymmetric triplets.

Figure 2.13 shows that within easy-8 triplets, those with the correct referent towards the low-density regions experienced a speed up post-exposure, consistent with density further stretching the psychological space between the query and the incorrect referent, making the choice easier. On the contrary, those asymmetric triplets with the correct referent towards the high-density side experienced neither a speed up nor a slow down. Given the overall post-pre acceleration in RTs (Figure 2.11), a lack of speed up in this case can be considered as slowing down, consistent with density stretching the space between the query and the referent in the dense region, making the choice harder.

We did not observe the same result within the easy-16 triplets. Given that for these triplets, one referent was much closer to the query than another, any effects of density on RT could have been drowned out by ease of detecting the right referent.

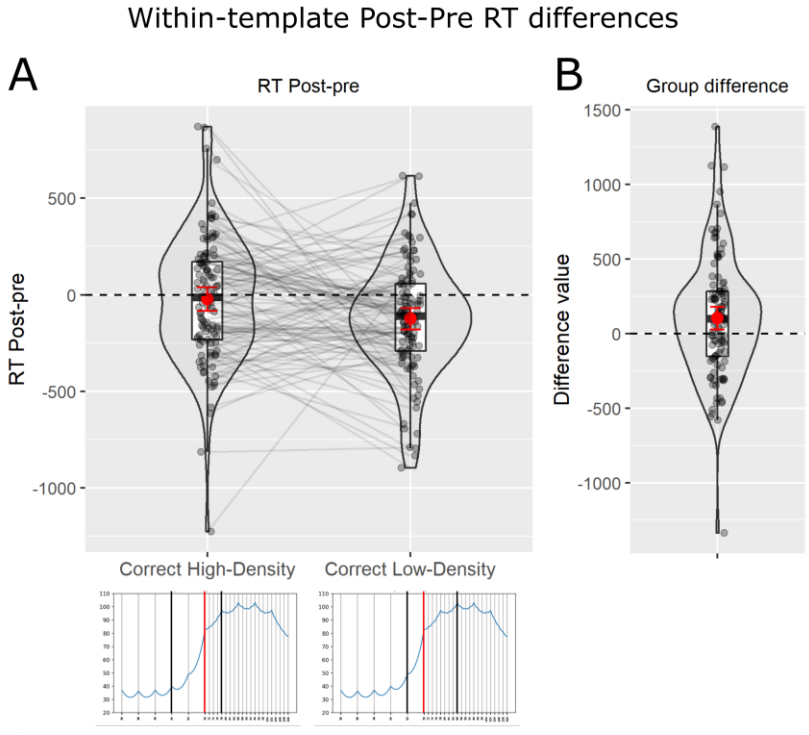


Figure 2.13: Post-pre RT for asymmetric triplets.

(A) Post-pre RT difference for asymmetric triplets that either have the "correct" referent in the high-density or the low-density region. The two insets below the plot show example triplet coordinates in corresponding density plots. (B) The difference

scores between the two types of asymmetric triplets. Each dot is a participant. Red dots signify group-level means. Error bars are 95% CIs.

2.3.2.4 Boundary effects

We found strong boundary effects in our stimulus space pre- as well as post-exposure. For each participant, we took symmetric triplets and calculated the probability of choosing the referent with the higher value on our curvature dimension. Without any bias, this probability should be roughly 0.5. As shown on Figure 2.14-A, this probability was significantly lower than 0.5 for those symmetric triplets that were in the convex part of the space, i.e. for which one of the referents was the leftmost boundary stimulus. For symmetric triplets in the middle region, the probability was indistinguishable from 0.5, whereas for those in the concave part it was significantly higher than 0.5. This effect was pronounced even for Block1 of pre-exposure (Figure 2.14-B). A likely source of this strong bias to choose boundary stimuli are the practice trials, where exemplars towards the boundaries of our 1D space were overrepresented (Figure 2.14-C).

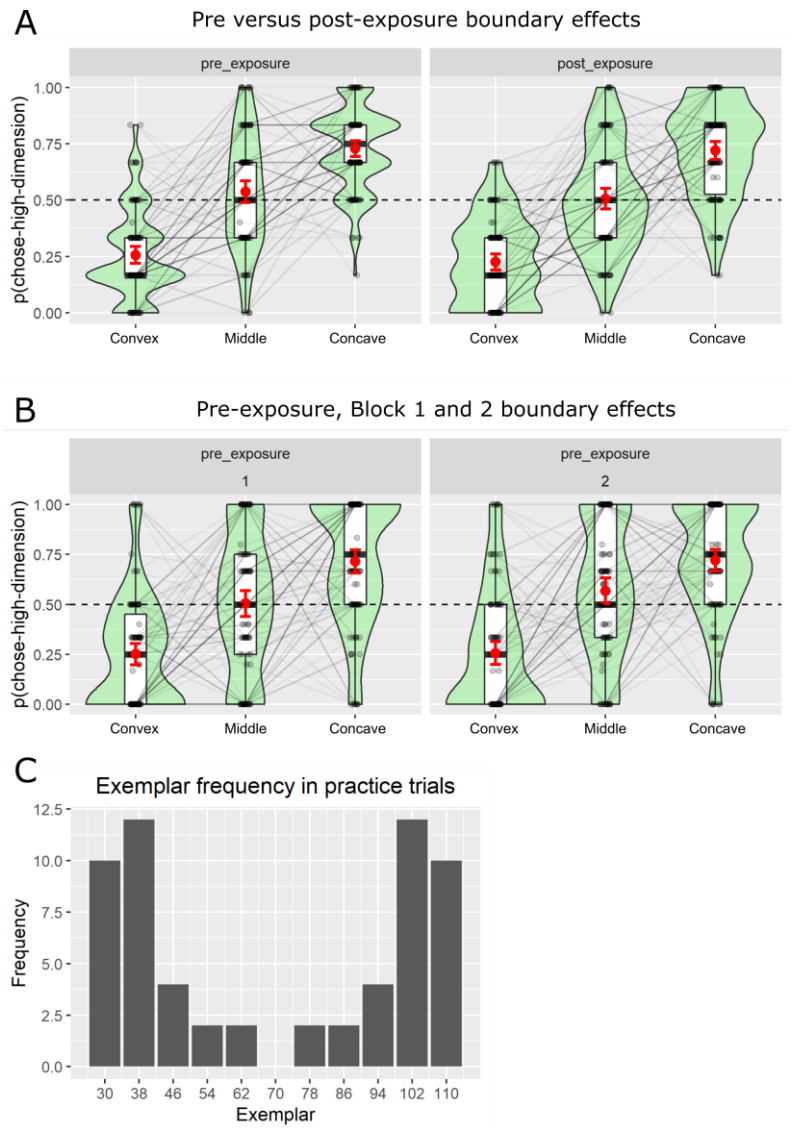


Figure 2.14: Boundary effects for symmetric triplets.

(A) Probability of choosing the referent towards the right side of our 1D stimulus space varied widely depending on the triplet location, demonstrating a strong bias for choosing the boundary stimulus as more similar to the query. (B) The boundary effect within pre-exposure blocks 1 and 2. The boundary effect was present already in Block 1. Each dot is a participant. Red dots signify group-level means. Error bars are 95% CIs. (C) Frequency distribution of exemplars shown during practice trials.

2.3.2.5 Summary of triplet matching task results

Overall, our main confirmatory analysis of middle symmetric triplets did not reveal an effect of density, as the referents in the low-density region were not chosen more frequently post- vs pre-exposure (Figure 2.10). Analysis of reaction times for middle symmetric triplets revealed a generic training effect (Figure 2.11), but RTs did not differ for middle triplets vs triplets from other regions (Figure 2.12), again indicating no effects

of density. The only exploratory analysis providing evidence in line with density effects concerned the asymmetric triplets, where we found that some asymmetric triplets with the "correct" referent towards the high-density regions did not show a post-pre RT speed-up compared to asymmetric triplets with the "correct" referent towards low-density regions. This would be expected if our exposure manipulation stretched the psychological space in dense regions, making the "correct" referent look further away than before, and resulting in a slowing down of judgments on such trials. This might indicate that our density manipulation created a linear density gradient across the whole distance of our 1D space, instead of a sharp shift as depicted in Figure 2.7. However, why such a gradient would not have influenced our main dependent variable of p(chose-low-density) in middle symmetric triplets remains unclear.

2.3.3 Results: the same-different task

A psychometric curve showing confusability as a function of the distance between back-to-back presented stimuli is shown in supplementary Supplementary Figure 8.1 in the Appendix.

2.3.3.1 Density does not influence confusability

A different way to test if density affected confusability is to compare items from high and low-density regions in terms of the probability of responding "same" on true "same" trials versus on true "different" trials, expecting better performance in high-density regions. Figure 2.15 shows plots for these two dependent variables calculated for the last two blocks of the same-different task, separately for stimuli in the high vs low density regions.

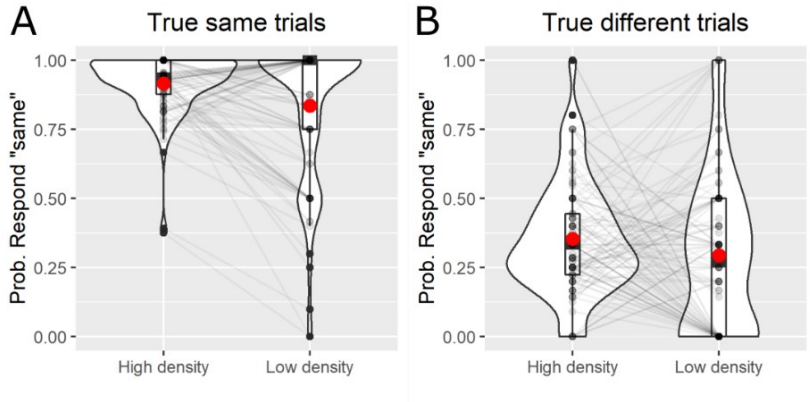


Figure 2.15: The "same" response bias during the same-different task.
Blocks 4-5 probability of responding "same" on either the "same" trials (A) or "different" trials (B) showed an overall bias to respond "same". Each dot is a participant. Red dots signify group-level means.

For the trials with items from the high-density region, we see an overall "same" response bias, regardless of the trial-type ("same" or "different"). A likely explanation is that the participants detected the presence of more exemplars from the high-density region, which had smaller distances between each other in the generative space, and therefore adopted a strategy to respond "same" more often. Thus, we do not find any effects of density on confusability either, apart from an overall response bias.

2.4 Discussion

2.4.1 Summary of the results

In this chapter, we tested the distance-density model proposed by Krumhansl (1978) as an answer to Tversky's (1977) challenges to classical geometric models of knowledge representation. In light of violations of several axioms, Tversky proposed that geometric models cannot be veridical algorithmic-level descriptions of conceptual representations, and proposed a feature-based representation along with a formal contrast model for calculating similarities. Along with the "focusing hypothesis", which specifies that more attention is given to the subject than the referent term during directional similarity statements, the contrast model accounts for asymmetric similarity judgments. Krumhansl (1987) proposed that augmenting classical geometric models by incorporation of local item density could account for axiom violations, without any need to discard geometric models. We tested the most basic prediction of such distance-density model: that increases in density should lead to decreases in similarity. Utilizing a novel 1D stimulus space and a within-participant manipulation of density, we did not obtain evidence that density impacted similarity.

We began with a norming study to validate our stimulus space and ensure that distances between exemplars in the generative space were linearly related to their distances in psychological space. By obtaining judgments on a triplet matching task, we estimated psychological embeddings using the PsiZ model (Roads & Mozer, 2019), confirming a linear relationship. This allowed us to make predictions for our main experiment: that after density manipulation, the probability of choosing the referent towards the high-density region should *decrease* as density stretches the space and this referent is perceived as further away. We manipulated density within-participants by exposing them to four times more stimuli from one part of the space during a same-different task, and then comparing similarity judgments for triplets post- versus pre- this exposure. However, we

did not find evidence of change in choice probabilities for the triplets located at the boundary of the density change, arguing against the distance-density model. Exploratory analysis of reaction times showed an overall speed-up post- vs pre-exposure, indicative of generic training effects. This speed-up did not differ for triplets in the middle region (where the density gradient might have made one referent look more distant than another, further decreasing the RT) compared to triplets from high-density and low-density regions. This further argued for a lack of density effects. Asymmetric triplets with the "correct" referent towards the high-density region did show a larger speed-up in reaction times compared to those with the "correct" referent towards the low-density region. Although reaction times can be a more sensitive measure than choice probabilities, and this result is in line with the distance-density model, this was an exploratory analysis that needs to be confirmed by follow-up experiments.

The same-different exposure task gave us another measure of similarity, through calculation of confusability. If density stretches the psychological space, accuracy should be higher on items from high-density regions towards final blocks of the task. However, we found an overall "same" response bias for items in the high-density region compared to items in the low-density region, without a specific increase in accuracy, arguing against any effects of density.

Finally, we found that when two referents were equally far in physical space, but one was a boundary stimulus at the end of our 1D distribution, participants were strongly biased towards the boundary stimulus. This bias was pronounced pre- and post-exposure, and was even present in Block 1 of the pre-exposure similarity task. Krumhansl (1978) argued that boundary stimuli are in a less dense region, which could explain some results in the literature, such as the finding that self-similarities for boundary items tend to be higher. Similar logic could explain our results. However, given that boundary effects are found across nearly every stimulus domain, and have been explained by more the general principle of distinctiveness (Murdock, 1960), as well as the observation that our practice trials were biased to over-represent boundary items, we cannot confidently point to density as the mechanism behind our boundary effects.

2.4.2 Limitations of the current study

It is possible that our null findings were due to several shortcomings of our paradigm. First, our triplet matching task might not have been sensitive enough to detect changes in similarity judgments. Every triplet was repeated only twice pre- and post-exposure task.

It is possible that a larger number of repetitions could provide a more sensitive assessment of any changes in probabilities of choosing a particular referent. Second, our density manipulation might not have been strong enough to exert influence on the underlying similarity structure, or the effects might take longer than the duration of our exposure task to manifest. Future paradigms could employ larger density changes or longer exposure tasks. Finally, despite powering our study to support H1 or H0, we did not reach our pre-defined threshold for BF_{01} after having run the maximum number of participants. Thus, we might have been underpowered to detect a small effect of density on similarity.

2.4.3 Relation to prior literature

Although the distance-density model has often been mentioned as a solution to challenges raised by Tversky (Markman, 2012; Nosofsky, 1992b), few prior experiments have directly tested its predictions with an efficient paradigm. Krumhansl (1987) reanalysed and reinterpreted data from previous similarity or discrimination tasks, arguing that they supported the predictions of her model. However, as pointed out by Corter (1987), either density was often confounded with other variables in those paradigms, or density was estimated using the primary discrimination data, making it circular to examine effects of density on the same data.

Corter (1987) directly tested the predictions of the distance-density model across several experiments, using either direct pairwise similarity judgments on ellipses, faces and letter-like figures, or a discrimination task on letters and letter-like figures. Density was manipulated in a between-participant manner by adding neighbours to certain target stimuli. Corter found no evidence of effects of density on similarity. It is possible, however, that due to weak density manipulation (addition of 3 exemplars in the neighbourhood), a passive exposure task (asking participants to simply flip through a booklet containing exemplar images) and a between-participant design with few participants per group, Corter was underpowered to detect an effect of density.

Instead of presenting neighbours, one could manipulate density by increasing the frequency of presentation of items. Although not discussing it in terms of density, Polk et al. (2002) took this approach of frequency manipulation, and using a within-participant, pre-post design, found that presenting some colour patches more often than others in an orthogonal size discrimination task caused asymmetries in a later directional colour similarity judgment task. Thus, without any change in the number of features or direct perceptual training, a simple frequency manipulation could lead to asymmetries.

Although the authors did not consider this, higher presentation frequency could also lead to higher density, supporting the distance-density model. However, we have argued that because presentation frequency can change stimulus saliency and saliency-induced asymmetries can be explained by Tversky's contrast model as well, the proper manipulation to disentangle predictions from the distance-density and the contrast model is to manipulate neighbourhood density through introduction of novel neighbouring exemplars.

One way in which exposure to neighbours could stretch the psychological space is through perceptual training, whereby participants become better attuned to fine-grained stimulus differences, which should lead to lower similarity judgments. In our paradigm, the participants did have to discriminate nearby stimuli within the high-density region. Despite this, we did not find any effects of this perceptual training on similarity judgments. One could argue, however, that longer training is necessary to elicit results. In a particularly relevant study, Collins and Behrmann (2020) examined the effects of multi-day discrimination training on similarity judgments. Across 20 days, the participants performed sample-to-match tasks involving specific stimuli drawn either from a face database or a set of artificial object stimuli (called UFOs). Crucially, they performed pair-wise similarity ratings of all the stimuli before, during and after the training. The authors found a global decrease in similarities across days, but found that specific exemplars used during the training task had the largest drop in similarity, becoming most distinct from other stimuli. Notably, training involved the same exemplars used in the first similarity task, with no new neighbour stimuli introduced. Thus, this stretching of psychological space could owe to simple frequency effects a la Polk et al. (2002), perceptual training, or both. It is important to note that although effects of perceptual training versus exemplar density would be comparable in terms of stretching the psychological space, perceptual training would not predict asymmetries in similarity judgments, unlike Krumhansl's proposal.

2.4.4 Future directions

In summary, the evidence to date (Corter, 1987; this chapter) does not support the predictions of the distance-density model, namely that increases in local neighbourhood can lead to changes in similarity, at least on short time-scales. This is despite findings that increases in saliency, as achieved by manipulation of frequency of presentation (Polk et al. 2002) or perceptual training (Collins and Behrmann 2020), have shown impacts on

similarity judgments. Further experiments could use longer paradigms with various exposure tasks to better differentiate the effects of saliency, perceptual training, and neighbourhood density on both (i) stretching of the psychological space and (ii) induction of asymmetries in similarity judgments.

Although the distance-density model has not been empirically supported so far, it is not the only augmented geometric model that could account for asymmetric similarity judgments. As discussed in the introductory Chapter 1 of this thesis, a contextualised geometric model where attention dynamically selects and weighs certain features during similarity calculations (see Equation 1.5) could in principle account for asymmetric similarity judgments (Decock & Douven, 2011; Gärdenfors, 2000). Such incorporations of attentional modulations begs the question for the origins of attentional shifts themselves. One explanation is that a larger emphasis comes to the subject term of the asymmetric similarity judgment compared to its referent term simply due to its temporal primacy. This could result in activation of different dimensions and re-weighing of distances along activated dimensions when the subject and referents items swap places. Thus, Tversky's documentation of symmetry violations, along with inability of Krumhansl's model to predict changes in similarities due to density, are insufficient to fully refute geometric models. Future experiments should systematically examine the validity of attention-weighted similarity calculations and specifically their ability to explain asymmetric judgments in the literature. In the next chapter, we test various two-dimensional stimulus spaces for adherence to geometric requirements and consider whether attention-weighted geometric models could explain our data.

3 The Triangle Inequality and Segmental Additivity

3.1 Introduction

In the previous chapter, we used a 1-dimensional novel stimulus space to test predictions of a distance-density model which tried to account for violations of the symmetry axiom. In the current chapter, we use 2-dimensional stimulus spaces to characterize adherence to two further requirements for geometric representations: the triangle inequality and segmental additivity.

3.1.1 The Triangle Inequality and Segmental Additivity

As discussed in the introductory chapter, segmental additivity and the triangle inequality capture basic intuitions underlying geometric spaces and map-building in general. Segmental additivity requires unidimensional additivity, i.e. that distances within any dimension should be additive, such that for any three points *a*, *b* and *c* lying on a segment, $D(a,c) = D(a,b) + D(b,c)$ (Figure 3.1-A). The triangle inequality applies to multiple dimensions, requiring inter-dimensional additivity or sub-additivity, and states that the shortest distance between any two points *a* and *c* must be a direct path, and that an indirect one passing through a third point *d* cannot be shorter: $D(a,c) \leq D(a,d) + D(d,c)$.

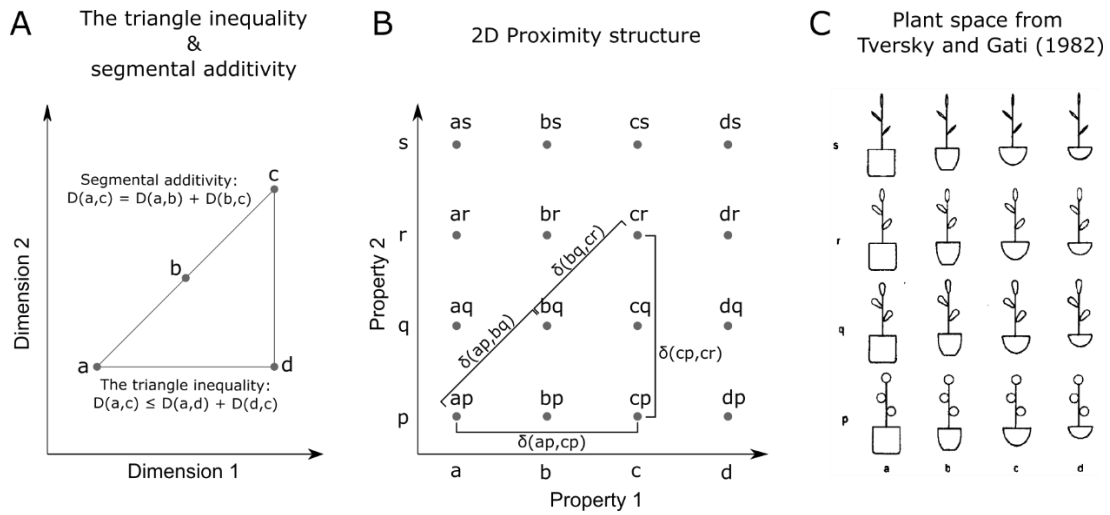


Figure 3.1: The triangle inequality, segmental additivity, and 2D stimulus spaces. (A) In a 2-dimensional space, segmental additivity requires that two segments (a,b) and (b,c) lying on the same path be additive, i.e. the sum of their distance must equal the distance between the endpoints (a,c). The triangle inequality requires that for any two points *a* and *c*, the distance between them cannot be larger than the distance through a point *d* which does not lie between *a* and *c*. (B) The two-dimensional proximity structure described by Tversky and Gati (1982). Each dimension consists of a set of coordinates A = {a,b,c,d} and P = {p,q,r,s}, with A x P being the product set consisting of all pairs ap, bq, bp, etc. Psychological distance between two points ap and bq is *δ(ap,bq)* which denotes an ordinal measure of dissimilarity. (C) The two-dimensional "plant space" tested by Tversky and Gati (1982). Plants varied by two qualitative attributes: shape of the pot and curvature of the leaf. Panel (C) adapted from Tversky and Gati (1982). Copyright © 2023 by American Psychological Association. Reproduced with permission. No further reproduction or distribution is permitted without written permission from the American Psychological Association.

Clearly, a 2-dimensional Euclidean physical space adheres to these requirements. To understand how one could test these axioms in psychological spaces using similarity judgments, we must consider (i) how physical distances along a single dimension translate to psychological unidimensional distances, and how (ii) psychological unidimensional distances get combined to form multi-dimensional distances.

As discussed in Chapter 2, much of the previous literature has characterized the relationship between physical and psychological spaces with an exponential function (e.g.

Nosofsky, 1985; Shepard, 1958). Tversky and Gati (1982) considered an alternative option of a power-law relationship:

Equation 3.1: $$\boldsymbol{\delta = D^{1/\lambda}}$$

Where D is the physical distance and δ the psychological dissimilarity. The advantage of this formulation is that it can accommodate a linear relationship between physical and psychological spaces as a special case, i.e. when $\lambda = 1$. Thus, in this chapter we consider that, along any single dimension, psychological perceived dissimilarities can be derived by a power-law transform of physical distances, with the λ as a free parameter to estimate.

When combining distances across dimensions in physical space, we can refer to the power metric formula from Equation 1.1 of Chapter 1, and simplify it for a two-dimensional case as:

Equation 3.2: $$\boldsymbol{D(a,c) = [D(a,d)^{\gamma} + D(d,c)^{\gamma}]^{1/\gamma}}$$

where a, c, and d are points in a two-dimensional space (Figure 3.1-A) and γ is the "Minkowski metric" specifying which distance metric to use.

Thus, putting Equation 3.1 and 3.2 together, we first translate unidimensional physical distances to psychological ones with the λ parameter, and then combine them into a multidimensional one using the γ Minkowski metric:

Equation 3.3: $$\boldsymbol{\delta(a,c) = \left[D(a,d)^{\frac{\gamma}{\lambda}} + D(d,c)^{\frac{\gamma}{\lambda}}\right]^{1/\gamma}}$$

To satisfy segmental additivity, such a model requires a linear mapping between unidimensional physical and psychological distances, i.e. $\lambda = 1$. To satisfy the triangle inequality, whether an indirect path is shorter than a direct one depends on γ, and it can be shown that with $\gamma < 1$, the diagonal of a right-angled triangle becomes larger than the sum of sides. Thus, the two requirements of segmental additivity and the triangle inequality set boundary conditions for the two parameters governing geometric representation of stimuli in psychological spaces: $\lambda = 1$ and $\gamma \geq 1$.

These parameters can be estimated by participants rating similarities of two of more stimuli, which are usually conducted with a Likert-style rating scale. However, this only provides an ordinal rather than continuous (interval) measurement. With ordinal data, one cannot directly test segmental additivity or the triangle inequality. To circumvent this, Tversky and Gati (1982) developed a novel method to test for the triangle inequality using ordinal measures.

3.1.2 Ordinal tests of the triangle inequality

Tversky and Gati (1982) considered points arranged in a 2-dimensional proximity structure, with each dimension consisting of a set of coordinates {a,b,c,d} and {p,q,r,s} (Figure 3.1-B). Considering a quadruplet of stimuli {ap,bq,cr,cp}, the triangle inequality is satisfied if the *centre path* between ap and cr that passes through bq is shorter than the *corner path* that passes through cp. In an ordinal sense, the corner path exceeds the centre path (so the triangle inequality is satisfied) whenever the unidimensional distances are larger than the two-dimensional ones:

Equation 3.4:

$$\delta(ap, cp) \geq \delta(ap, bq) \quad \text{and} \quad \delta(cp, cr) \geq \delta(bq, cr)$$

or

$$\delta(ap, cp) \geq \delta(bq, cr) \quad \text{and} \quad \delta(cp, cr) \geq \delta(ap, bq)$$

provided that at least one of the above inequalities is strict. If the opposite pattern of inequalities holds, then the centre path exceeds the corner path, violating the triangle inequality. In all remaining cases, ordinal data do not provide sufficient information for testing the triangle inequality.

Given this methodological insight, the authors reviewed existing studies (Burns et al., 1978) and conducted new experiments using pair-wise dissimilarity ratings on stimuli drawn from various 2D conceptual and perceptual spaces:

1. Plants varying on two qualitative attributes: form of the pot and elongation of leaves (Figure 3.1-C)
2. Plants varying on quantitative and qualitative attributes: size of the plant and elongation of leaves.
3. Students described as varying on conceptual qualitative attributes: major of study and political affiliation.
4. Dial-like figures varying in quantitative attributes of circle size and angle of the radial line (adapted from Shepard, 1964).
5. Squares varying in quantitative and qualitative attributes of size and brightness.
6. Colour patches varying in qualitative attributes of hue and chroma.

The first four stimulus spaces above have psychologically separable dimensions while the sixth entails psychologically integral dimensions (see Chapter 1 section 1.1 for definitions). Prior work had shown that the dimensions of the fifth stimulus space (squares) are sometimes perceived as separable and sometimes as integral (Burns et al., 1978).

For the first four studies, analysis of pair-wise ratings showed consistent violations of ordinal triangle inequality. The fifth study led to a bimodal distribution of satisfaction and violation, while the sixth study showed no violations. Thus, whether the triangle inequality is violated seems to depend on the nature of the stimuli being rated.

To corroborate these results, the authors conducted several tests for the triangle inequality which treated similarity ratings as interval data instead. This allowed them to directly estimate the γ Minkowski exponent of the distance function (in Equation 1.1), which they found to be less than 1 for the first five of the six studies, which indicated violations of the triangle inequality. While pointing out the inappropriateness of running such tests on pair-wise ordinal ratings, Tversky and Gati argued that these results supported the ordinal tests in the conclusion that such psychologically separable two-dimensional stimuli are not accurately described using a geometric representational model.

Summarizing their results, the authors argued that while geometric theories might be useful to think about and visualize similarity data, they cannot be veridical algorithmic-level theories of concept representation, since they violate fundamental requirements of metric axioms.

3.1.3 The current experiment

In this chapter, we tested the geometric properties of various 2D stimulus spaces that were adapted from recent neuroimaging literature (Constantinescu et al., 2016; Theves et al., 2019). This literature indirectly supports geometric theories by showing a parallel between neural computations underlying spatial navigation and "conceptual navigation" (see Chapter 1 section 1.2). Across six different stimulus spaces employing various types of dimensions, participants provided similarity ratings for every possible pair of exemplars. In brief, we show that some of our 2D spaces violate ordinal triangle inequality while others satisfy it, likely explained by differences in the nature of dimensions. In contrast, when estimating γ directly, none of the groups of stimuli had a γ value of less than 1, contrary to expectations based on Tversky and Gati's work (1982).

To understand this discrepancy between ordinal (the ordinal triangle inequality test) and interval (γ estimation) results, we ran simulations of ideal observer data generated using Equation 3.3 with a range of values for λ and γ. The simulations showed that violations of ordinal triangle inequality can be due either to $\gamma < 1$ or $\lambda > 1$. Given that our γ estimates were all above unity, this would indicate that $\lambda > 1$ implying violations of segmental additivity stemming from non-linear mapping between physical and psychological distances. Importantly, this pattern would not be compatible with either classical or augmented attention-weighted geometric models. However, our simulations also revealed inherent noise in our γ estimation procedure, due to which the attention-weighted geometric model cannot be fully refuted.

3.2 Experiment

3.2.1 Methods

3.2.1.1 Participants

134 healthy young adult participants were recruited (72 females) from the prolific.co platform, aged 19-42 (M = 29.115, SD = 6.02), and paid £6/hour for their time, according to the Cambridge Psychology Research Ethics Committee protocol PRE.2020.018. Of these, 73 (34 females, 54.48% of those recruited) aged 19-42 (M = 28.944, SD = 6.143) passed the final quality and performance checks (see below) to be included in the data analysis.

3.2.1.2 Stimuli

We created six distinct 2-dimensional stimulus spaces employing different types of dimensions. All stimuli were designed in Inkscape (Inkscape 1.1, https://inkscape.org) and the Psychtoolbox package (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) in Matlab R2020a (www.mathworks.com). For each of the spaces, four levels were chosen for each dimension, resulting in 16 unique exemplars. See Figure 3.3 for 2D depiction of each stimulus space. The six spaces were grouped into three groups depending on whether the stimuli were naturalistic or artificial and whether their dimensions were quantitative, qualitative or a mix of the two. All the stimulus spaces were designed to have psychologically separable dimensions, as Tversky and Gati (1982) found violations of the triangle inequality in such spaces.

Group 1: naturalistic stimuli varying on quantitative dimensions:

- **Birds defined by neck and leg length:** Adapted from Constantinescu et al. (2016), these stimuli were birds with varying lengths of neck and legs. Neck and leg lengths were sampled linearly for the 4 levels with lengths of 127, 185, 243, and 302 pixels.
- **Birds defined by beak and tail length:** Similar types of birds but with varying lengths of beak and tail, varying across 113, 164, 216, and 268 pixels.

Group 2: naturalistic stimuli varying on qualitative shape dimensions:

- **Plants defined by pot and leaf shape:** Based loosely on the 2D plant space used by Tversky and Gati (1982, Figure 6), these stimuli were plants varying in the shape of their leaves and of their pots. The shape of the leaf was varied by changing its width with the following four levels 35, 69, 102, and 135 pixels. The shape of the pots was manipulated by changing the width of its top segment, with the four levels being 119, 194, 269, and 345 pixels.
- **Lamps defined by base and shade shape:** lamps varying in the shape of the base and width of the shade. The pixel heights for the base levels were 29, 89, 150, and 210 pixels, while for the shade width were 149, 224, 299 and 375 pixels.

Group 3: artificial stimuli varying on qualitative and quantitative dimensions:

- **“Squircles” defined by the opacity of the square and the size of the circle:** an artificial stimulus adapted from Theves et al. (2019) that we called “Squircles”, consisting of squares varying in their opacity and circles varying in their size. Circle radius levels were 72, 104, 136, and 168 pixels. The four opacity levels were determined by the experimenter such that each step resulted in a roughly equal change in perceived opacity, resulting in the following values: 10%, 30%, 60%, 100%.
- **“Stripeys” defined by spatial frequency of the square and size of the circle:** The sixth stimulus space was an artificial stimulus called “Stripeys”, varying in the spatial frequency of the square and the size of the circle. Circle radius levels were 72, 104, 136, and 168. The four spatial frequency levels were determined by the experimenter such that each step resulted in a roughly equal change in perceived frequency, resulting in the following spatial frequency per pixel values: 0.125, 0.25, 0.5, 1.

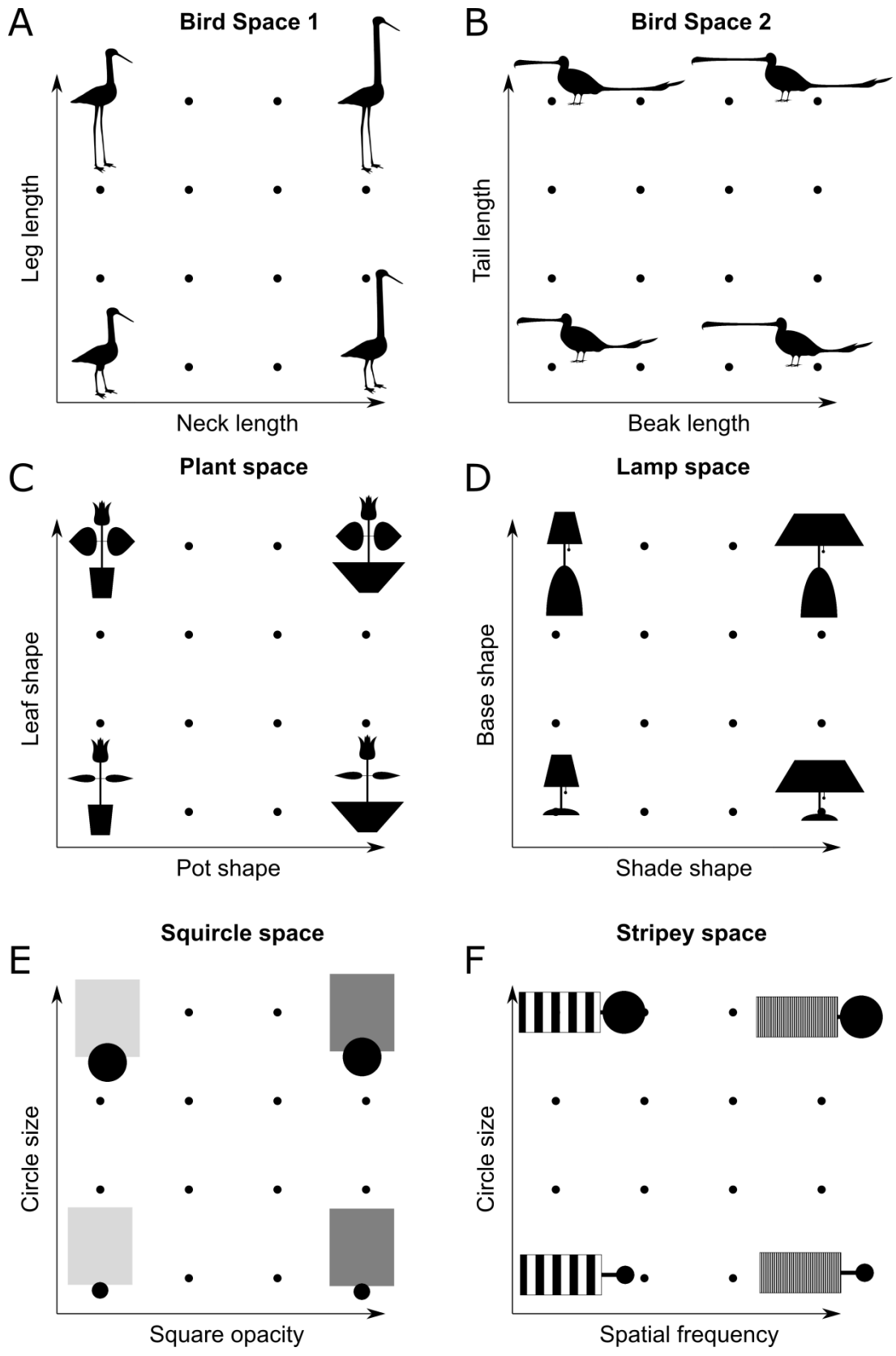


Figure 3.2: Two-dimensional stimulus spaces used in the experiment.

(A) Birds varying along quantitative dimensions of the length of their necks and legs. Stimuli adapted from Constantinescu et al. (2016). (B) Birds varying along quantitative dimensions of the length of their beaks and tails. (C) Plants varying

along qualitative dimensions of pot and leaf shapes. These stimuli were designed to resemble the plant space of Tversky and Gati (1982) shown in Figure 3.1-C. (D) Lamps varying along qualitative dimensions of shape of their shade and base parts. (E) "Squircle" stimuli consisting of a square varying along a qualitative dimension of opacity and a circle varying along a quantitative dimension of size. These stimuli were designed to resemble the 2D abstract space of Theves et al. (2019). (F) "Stripey" stimuli consisting of a square varying along a qualitative dimension of spatial frequency of stripes and a circle varying along a quantitative dimension of size.

3.2.1.3 Task design and procedure

3.2.1.3.1 Consent and Instructions

After consenting, participants were randomly assigned to one of the six stimulus conditions. The following is an example instructions text for the Stripeys stimulus group:

"In this experiment, you will be shown two pictures of Stripeys – artificial objects consisting of two shapes: a striped square and a circle. You will be asked to indicate on a 10-point scale how similar the two pictures look to you. For example, if the pictures of Stripeys are very different from one another, click on the lower number on the scale corresponding to low similarity. If the pictures are very similar, click on the higher number on the scale corresponding to high similarity. If the pictures look identical, then choose the highest number on the scale. In the same fashion, for all intermediate levels of similarity, use the intermediate values of the scale depending on your judged degree of similarity.

For the trials with pictures that are identical, we expect you to use the highest rating on the scale. For all other trials, there are no correct or incorrect answers. We are interested in your subjective impression of the degree of similarity, and different people are likely to have different impressions. Simply look at the two pictures for a short time, and click on the number that appears to correspond to the degree of similarity between them."

After the instructions, the participants were presented with all 16 exemplars of the concept all at once on the screen for at least 10 seconds (see example on Figure 3.4-B), for the purpose of familiarizing them with the range of variation between the exemplars. They were then informed that certain performance checks would be running throughout the experiment, failing which would result in being discontinued.

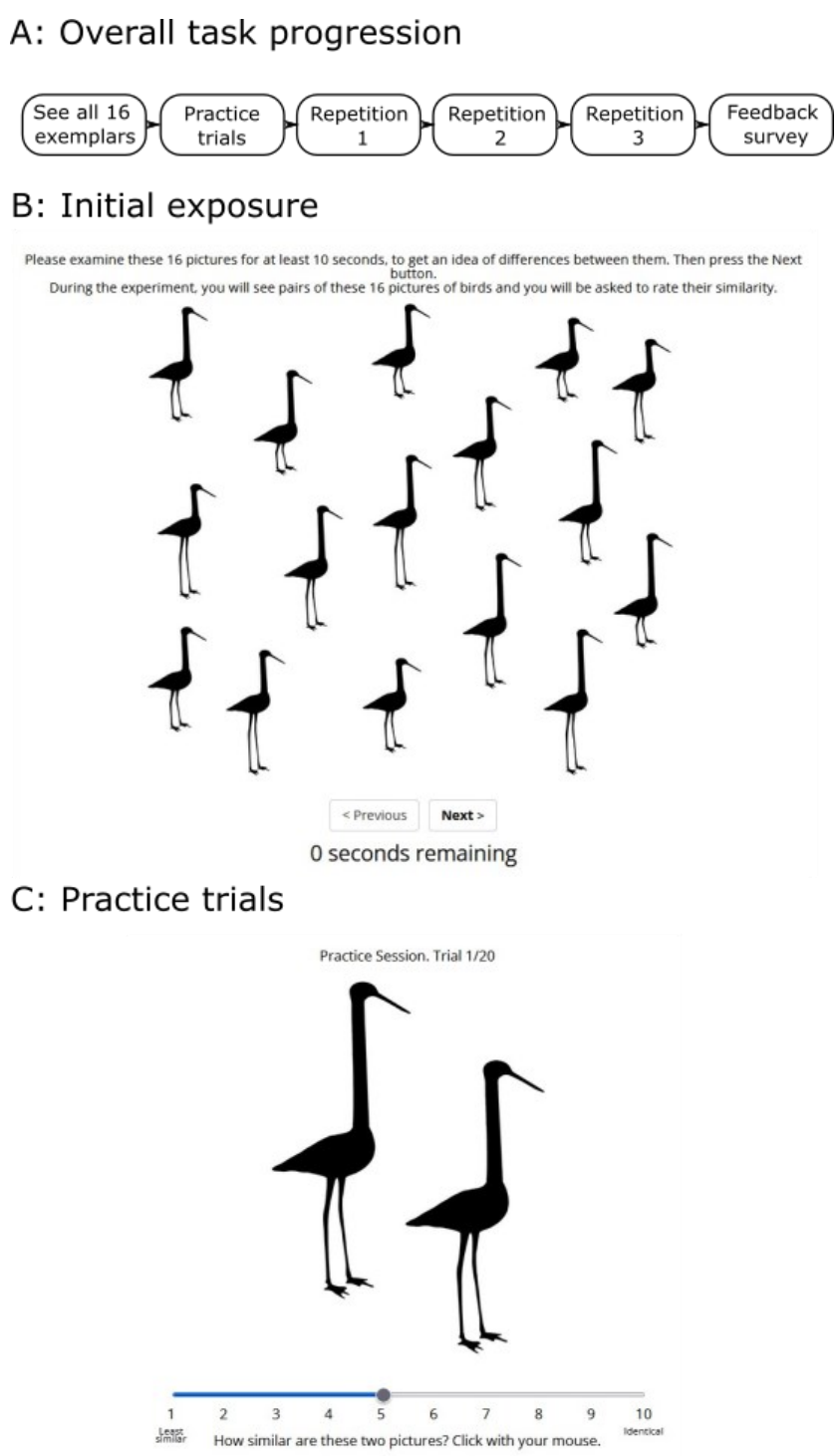


Figure 3.3: The pair-wise similarity ratings task.

(A) The overall structure of the experiment. Participants started by observing all 16 exemplars to familiarise with the range of variation in features. This was followed by 20 practice trials to familiarize with the trial structure. Afterwards, the main trials began, with 3 repetitions of all pairs of stimuli. Finally, a feedback survey and a debriefing were provided. (B) An example screenshot of initial exposure to all 16 exemplars of the birds from the neck:legs space. Participants had to observe the

stimuli for at least 10 seconds, after which they could advance. (C) An example practice trial presenting two birds from the neck:legs space.

3.2.1.3.2 Practice trials

The practice trials consisted of 20 trials including all 16 stimuli (Figure 3.4-C). At least 2 trials contained identical pairs, serving as catch trials. On each trial, the participants saw two exemplars displayed next to each other with either a vertical or horizontal offset to make the comparison harder. The participants had 30 seconds to respond on a similarity scale of 1 (least similar) to 10 (identical). The next trial began after an ITI of 500ms.

If the participants failed to respond with an 8 or above on any of the two catch trials containing identical exemplars, they were shown the two exemplars again, informed that they should have responded with the highest value, and redirected to re-read the instructions and re-do the practice trials. If the participants failed to respond accordingly the second time, they were discontinued from the study. Any missed trials were also repeated.

3.2.1.3.3 Experiment trials

Total of 128 pairs of exemplars were used for the similarity task, consisting of 120 unique pairs and additional 8 pairs of identical exemplars. These 128 pairs were repeated three times, with each repetition occurring across two blocks of 64 trials. For each repetition of a trial, the order of exemplars on the screen was flipped (counterbalanced across participants).

Each block was followed by a mandatory 15 second break, while a 30 second break took place between repetitions. Trial presentation was the same as for the practice trials, except that each trial lasted for 10 seconds, with an ITI of 500ms. Any missed trials were repeated at the end of every block. If the participants still missed those trials the second time, they were discontinued from the study.

3.2.1.3.4 Debriefing

At the end of the experiment, the participants were asked some debriefing question about any task strategies or other feedback.

3.2.1.4 Quality checks

Only desktop/laptop users were allowed to participate (no mobile devices). Minimum screen size requirement was set to 700x750. After each task block, responses on the trials

of that block were checked for the following criteria, which were developed through piloting:

- Less than 40% missed trials.
- Less than 20% of the trials with RTs below 1000ms.
- Less than 85% of the trials with RTs below 1500ms.
- No one similarity response value occurred in more than 35% of the trials.
- No combination of 3 similarity responses comprised more than 80% of the trials.
- No one response was repeated more than 6 times consecutively.
- Using a sliding window of 10 trials, none of the windows contained eight or more of the same responses.
- No more than five repeats of three consecutive uniform responses.
- Each catch trial with identical exemplars received a response of "8" or higher. If not, those trials were repeated at the end of the block. If a response of "8" or above was still not given, the participant was discontinued.

Additionally, we excluded any participant whose mean reaction time was three standard deviations (SD) from the mean across participants, as well as those who indicated in their debriefing that they failed to follow instructions. Examples of failing to follow instructions included: not understanding the similarity judgment, failing to notice that the stimuli varied along one of the 2 dimensions, etc.

3.2.1.5 Data analysis

We used Matlab R2020a (www.mathworks.com) and R RStudio (http://www.rstudio.com/) with R statistical software (R Core Team, 2022) for data preprocessing and analysis.

The reported similarity values were transformed into distances using the following formula: distances = max(similarities) - similarities. Thus, a reported similarity values of 10 (identical) and 1 (least similar) got translated to a distance of 0 and 9, respectively. Following prior literature (Attneave, 1950), we assumed the participants used the first repetition to adjust to the range of variation in stimulus sizes. Thus, all analysis was performed on averaged similarity data across the last two repetitions.

3.2.1.5.1 Basic assumptions of a 2D proximity structure

Following Tversky and Gati (1982), we first checked whether our 2D spaces adhered to basic assumptions for a well-defined 2D proximity structure. A description of these assumptions and results are presented in the appendix. Briefly, the 2D spaces adhered to all assumptions except for transitivity of betweenness, due to which we restricted our analysis to smaller triangles and segments.

3.2.1.5.2 The ordinal tests of the triangle inequality

The triangle inequality can be tested with ordinal data by comparing if unidimensional distances exceed the two-dimensional ones. For each participant, we used Equation 3.4 to classify number triangles that satisfied or violated ordinal triangle inequality for every quadruplet stimuli of the form {ap,bq,cr,cp} in Figure 3.1-B. To determine participant-specific chance level, we permuted similarity ratings 10,000 times to obtain a null distribution of expected number of triangle inequality satisfactions.

To test for group differences, we used the non-parametric Kruskal-Wallis one-way analysis of variance (given the bounded and skewed distribution of values; see Figure 3.4), followed by multiple pair-wise Wilcoxon rank sum tests with Bonferroni correction for 3 tests.

3.2.1.5.3 γ estimation

Another way to test the triangle inequality is to directly estimate the exponent γ in the Minkowski distance. Tversky and Gati (1982) utilized three different methods for γ estimation, each relying on certain assumptions. Here, we used a different approach. We considered right-angled *corner triangles* in our 4x4 space, such as the triangles formed by triplet of stimuli {ap,bq,bp} or {ap,cr,cp}. Assuming that the relationship between the diagonal and the sides is governed by the power metric Equation 3.2, we used the *optimize* function in R to estimate the γ parameter by minimizing over the sum of squares difference between the diagonal and the sum of the unidimensional distances:

Equation 3.5: $$\min_{\gamma} \sum_{i}^{n} [\boldsymbol{\delta_i}(\boldsymbol{diagonal}) - (\boldsymbol{\delta_i}(\boldsymbol{side_1})^{\gamma} + \boldsymbol{\delta_i}(\boldsymbol{side_2})^{\gamma})^{(1/\gamma)}]^2$$

where *i* denotes a specific corner triangle and *n* denotes total number of such triangles.

Bounds for γ parameter were set between 0 and 10. We used a standard non-parametric outlier exclusion criterion based on the first and third quartiles (Q1 and Q3) and the interquartile range (IQR). Any resulting γ value outside the range of Q1-1.5*IQR to Q3+1.5*IQR was deemed as an outlier and excluded from the analysis.

Because of the violations of transitivity (see the appendix), out of 144 possible corner triangles, the γ estimation analysis was performed on 100 triangles that did not involve sides that spanned more than 2 levels on either dimension.

3.2.1.5.4 Ideal observer simulations

To determine the sensitivity and specificity of the ordinal tests of the triangle inequality, as well as the reliability of our γ estimation procedure, we generated pair-wise similarity ratings under multiple simulated ideal observer agents with different underlying λ and γ parameters that either satisfied or violated geometric axioms. For a given simulated agent, a response was generated using Equation 3.3 and mapped onto a 1-10 similarity rating scale, using the following four steps:

1. For every pair of stimuli, generate perceived psychological distances p_δ from physical distances D using Equation 3.1.
2. Transform distances into perceived similarities, $p_s = 1 - p_\delta$
3. Use a *nearest neighbour* approach to map p_s onto a 1-10 rating scale to obtain "reported similarity" values, r_s. Under this approach, the perceived similarity values are first normalized to 1-10. The lowest perceived similarity is mapped to the lowest scale value (1), the highest perceived similarity is mapped to the highest scale value (10), while intermediate values get mapped onto their closest neighbouring value between 1 and 10.
4. The resulting r_s values were finally transformed to dissimilarities r_δ to be used for ordinal triangle inequality tests and γ estimation procedures. The same formula was used as for real participants: $r_\delta = \max(r_s) - r_s$.

For each simulated ideal observer, we subjected the resulting distance values to ordinal triangle inequality tests and the γ estimation procedure.

3.2.2 Results

3.2.2.1 The ordinal triangle inequality analysis

For the 16 right-angled triangles where the diagonal consisted of two segments (see Figure 3.1-B), we classified for each participant how many satisfied, violated, or did not give conclusive data for assessing ordinal triangle inequality (Supplemental Figure 8.3). In Figure 3.4, we plot distributions of percentile values (relative to permuted null distributions) for satisfying and violating ordinal triangle inequality for each group.

Kruskal-Wallis Test for one-way analysis of variance on the percentile data of ordinal triangle inequality satisfaction showed that there was a statistically significant difference $\chi^2(2) = 30.86$, $p < 0.001$. Pairwise comparisons using Wilcoxon rank sum test revealed that all groups significantly differed from each other (Group 1 vs 2 $p = 0.014$, Group 1 vs 3 $p < 0.001$, Group 2 vs 3 $p < 0.001$, Bonferroni corrected). These results show that naturalistic bird stimuli of Group 1 with quantitative length dimensions did not violate ordinal triangle inequality, artificial non-verbalizable stimuli of Group 3 with quantitative and qualitative dimensions did violate it, while the naturalistic Group 2 stimuli with qualitative dimensions did not provide data to support either conclusion.

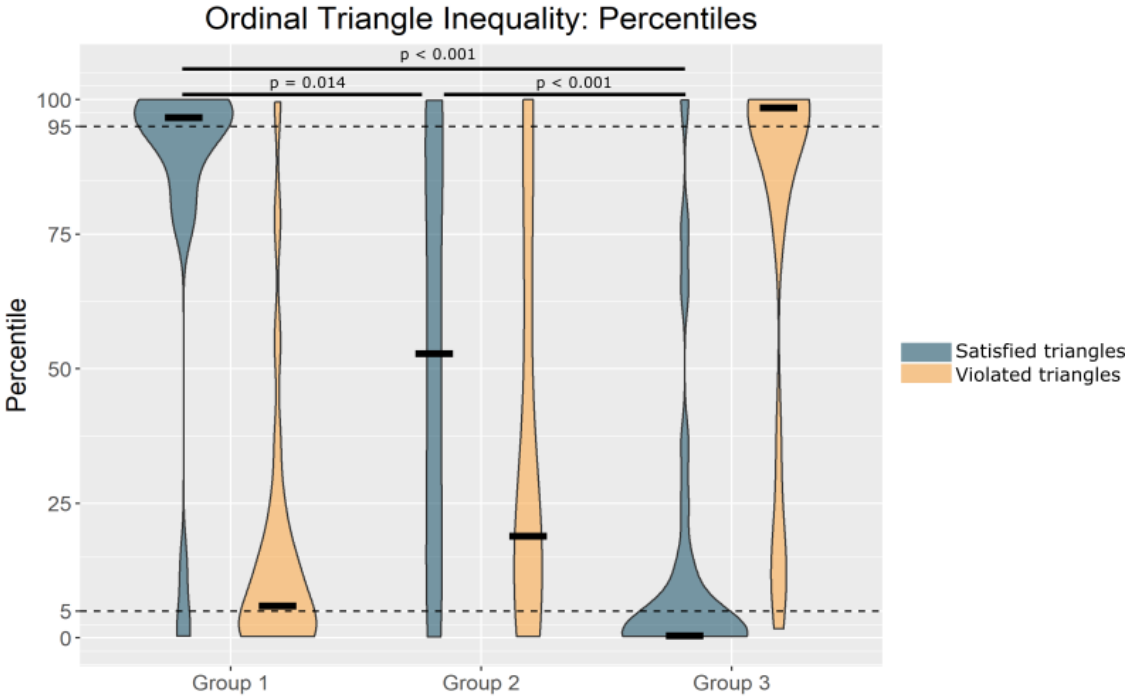


Figure 3.4: Percentile values for satisfying or violating ordinal triangle inequality relative to participant-specific permuted distributions.

Black bars show the median values for each group. For each group, blue distribution shows values for the number of satisfied triangles while orange distribution signifies percentile values for the number of violated triangles. Group 1 satisfied and did not violate ordinal triangle inequality (median percentile satisfied = 96.7, median percentile violated = 5.95). Group 2 data did not show clear patterns of satisfaction or violation (median percentile satisfied = 52.8, median percentile violated = 19). Group 3 did not satisfy and violated ordinal triangle inequality (median percentile satisfied = 0.38, median percentile violated = 98.5). Note that for each participant, these two values do not have to add up to 100, since the number of satisfied and

violated triangles do not have to add up to the total number of triangles (see Supplemental Figure 8.3 in the Appendix).

3.2.2.2 γ estimation

Tversky and Gati (1982) used various methods to estimate the γ exponent in the distance function (Equation 3.2), finding $\gamma < 1$ for five of their six stimulus spaces. Using our γ estimation method across corner triangles in the 4x4 space, we find that, contrary to Tversky and Gati, γ estimates for all three stimulus groups were above unity: Group 1 Mean = 2.21, 95% CI [1.78, 2.64], Group 2 Mean = 2.14, 95% CI [1.54, 2.74], Group 3 Mean = 1.42, 95% CI [1.14, 2.71] (Figure 3.5). Data for 5 out of 24 Group 1 participants and 3 out of 24 Group 2 participants were excluded due to extreme outlier status. All outliers were biased towards larger γ estimates ($\gamma > 5.6$).

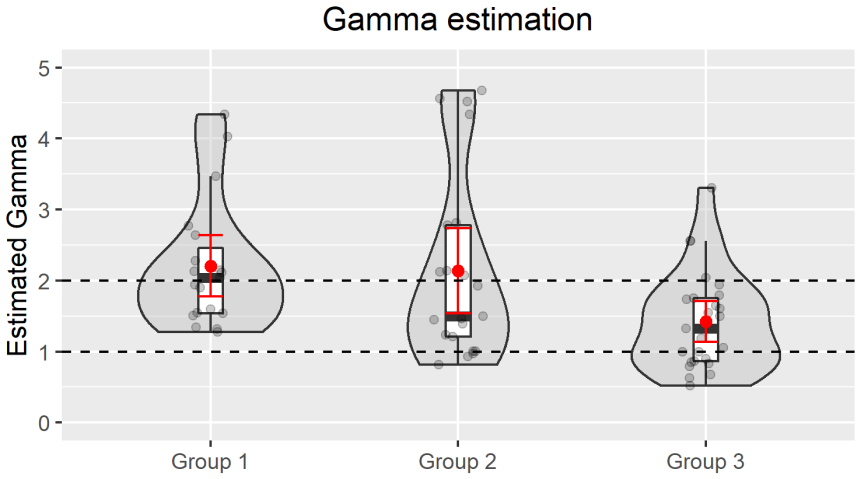


Figure 3.5: Minkowski parameter estimates for the three stimulus groups.
Mean and median estimates of each group were above unity. Each dot is a participant. Red dots signify group-level means. Error bars are 95% CIs. The two horizontal dashed lines at $\gamma = 1$ and $\gamma = 2$ are shown for reference and correspond to the city-block and Euclidean metrics, respectively.

To further illustrate and clarify the relationship between the direct path (i.e. the diagonal) and the corner path (i.e. sum of sides) in our corner triangles, Figure 3.6 below plots the relationship separately for each of our three stimulus groups (compare to an analogous Figure 7 of Tversky and Gati, 1982). Each dot represents data for a corner triangle (averaged over participants), with the x axis coordinate corresponding to the corner path distance and the y axis coordinate corresponding to the direct path distance. We can see that Group 1 and Group 2 triangles are sub-additive across dimensions, characteristic of

$\gamma \geq 1$, while Group 3 triangles are roughly additive across dimensions, characteristic of a city-block metric where diagonal distances are simply the sums of unidimensional distances. Note, however, that such additivity analysis as well as the γ estimation analysis rely on interval nature of measurements, whereas pair-wise ratings can only provide ordinal scale measures. Thus, these results should be interpreted with caution.

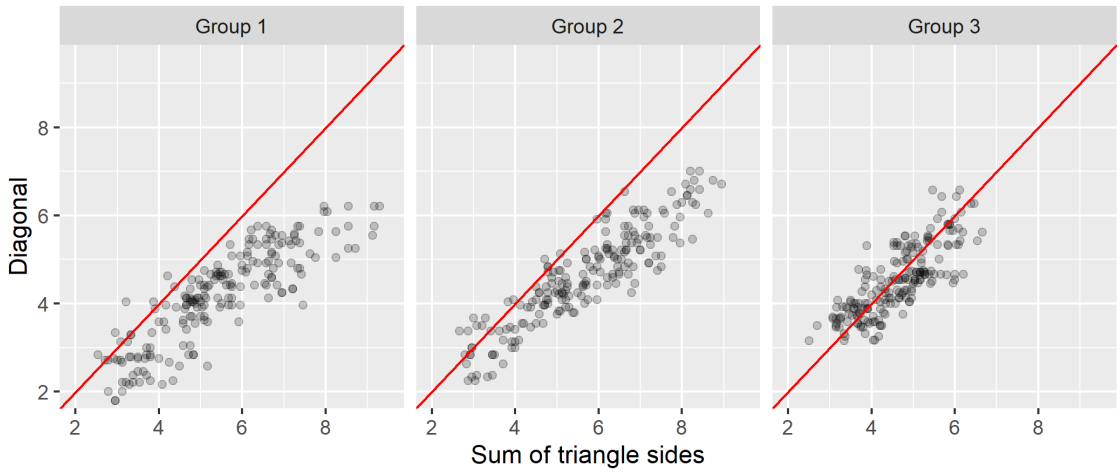


Figure 3.6: Additivity analysis for corner triangles.
Each dot represents a corner triangle, with data averaged over participants within the group. Red line is a 45 degree line. Dots lying above the red line violate the triangle inequality, implying $\gamma < 1$. Those on the red line show additivity, i.e. implying $\gamma = 1$. Dots below the red line indicate satisfaction of triangle inequality, implying $\gamma \geq 1$. Group 1 and Group 2 stimuli appear sub-additive across dimensions. Group 3 stimuli appear roughly additive $\gamma = 1$.

3.2.2.3 Ideal observer simulations

The ideal observer simulations allowed us to test to what extent the ordinal triangle inequality method and the γ estimation methods reflect the metric properties underlying the process generating similarity data.

3.2.2.3.1 Validating the ordinal triangle inequality test

For each simulated ideal observer with unique combination of λ and γ parameters, we tracked satisfaction/violation of ordinal triangle inequality for (i) the Euclidean distances D between generative physical coordinates of stimuli, (ii) the continuous perceived psychological distances p_δ, and (iii) the reported distance values after mapping of psychological values to ordinal rating scale r_δ (Figure 3.7, the three rows).

Figure 3.7: Ideal observer simulations for the ordinal triangle inequality test. Each column of panels shows results for simulations with a different λ value, showing results only for λ = 1, 1.1, 1.2 and 1.3 for brevity. The *x* axis within each panel shows different values of simulated γ. Each row of panels depicts ordinal triangle inequality tests applied to a different outcome variable: top panel shows results for distances in physical Euclidean space *D*, middle panel depicts results for perceived continuous dissimilarity values p_δ, bottom panel shows results for reported dissimilarity values r_δ after mapping onto a discrete 1-10 scale.

For Euclidean distances, the ordinal inequality outcome is fully governed by the γ parameter: it is violated if $\gamma < 1$, satisfied when $\gamma > 1$, and non-diagnostic when $\gamma = 1$.

For the continuous perceived psychological distance values p_δ, since λ and γ trade off of each other in Equation 3.3, the inflection point of violation/satisfaction corresponds to where $\lambda = \gamma$. Ordinal triangle inequality is satisfied whenever $\gamma > 1$ and $\gamma > \lambda$. In a hypothetical experiment with access to internal psychological continuous distance representations between stimuli, these simulations would help interpret empirical results. If the psychological data show violation of ordinal triangle inequality, we could conclude that either $\gamma < 1$ or $\lambda > 1$, either of which violate metric requirements. If, on the other hand, ordinal tests show satisfaction of triangle inequality, such data are compatible with metric models when $\lambda = 1$ and $\gamma > 1$, as well as non-metric ones when $\gamma > 1$ but $\lambda > 1$. Thus, in the latter case, no definitive inference can be made regarding the metric nature of the underlying generative process.

Finally, the bottom row of Figure 3.7 shows the simulation results for r_δ – the psychological distances mapped onto a discrete scale. We can see that going from perceived dissimilarities p_δ to reported dissimilarities r_δ involves loss of information, expressed by large parts of the parameter space where ordinal triangle inequality is not

diagnosable (black dotted lines). However, if empirically reported similarity values violate ordinal triangle inequality, such a result is compatible only with $\gamma < 1$ or $\lambda > 1$, allowing us to infer violation of metric requirements. On the flip side, as with the continuous perceived psychological distances, satisfaction of ordinal triangle inequality is compatible with both metric ($\lambda = 1$, $\gamma > 1$) and non-metric generative spaces ($\gamma > 1$ but $\lambda > 1$ too).

These simulations help interpret the empirical results of our experiment (Figure 3.4 and Figure 3.5): Since the artificial stimulus spaces (Group 3) violated ordinal triangle inequality, we can infer that the underlying psychological space is characterized by a non-metric model with either a γ less than 1, or λ that is larger than 1 (or both). The bird spaces (Group 1), on the other hand, despite satisfying ordinal triangle inequality, are compatible with both metric ($\gamma > 1$ and $\lambda = 1$), but also non-metric ($\lambda > 1$) spaces.

3.2.2.3.2 Validating the γ estimation method

Figure 3.8 below plots the estimated γ values for ideal observers characterized by a range of γ and λ values. We can see that, due to mapping of continuous psychological variables onto a discrete scale, the estimation procedure is noisy, sometimes underestimating and sometimes overestimating the true γ value. That this imprecision is caused by the mapping procedure is demonstrated by the fact that estimating γ directly on the internal psychological distance values, p_δ, perfectly recovers the true parameter (see Supplemental Figure 8.4 in the Appendix).

Importantly, the simulations showed that for a range of λ values (e.g., highlighted values of 2.0 and 2.2), our estimation procedure overestimates γ to be equal or more than 1, while the true γ is < 1. This occurs for γ in the range of 0.9 and 1. This complex interaction between the λ parameter and mis-estimation of γ makes our empirical estimates difficult to interpret. Specifically, it is possible that artificial stimuli of Group 3 have true γ value less than 1, even though our γ estimates were larger than 1. This has consequences for our data in terms of ruling out attention-weighted geometric models, as we elaborate on in the discussion.

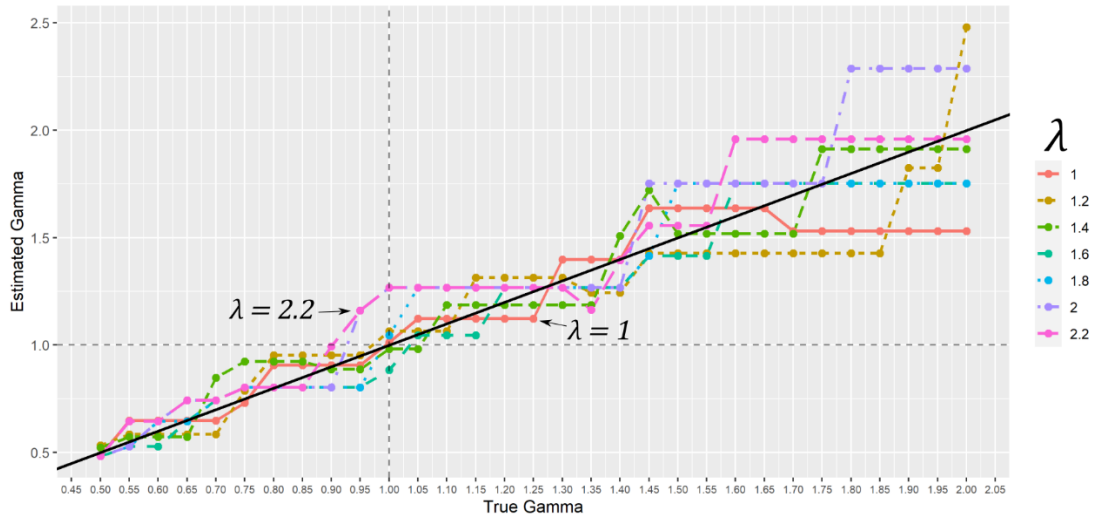


Figure 3.8: γ estimation procedure applied to simulated ideal observer data. For each simulated participant with underlying γ and λ values for Equation 3.3, we took their final reported distance values r_δ and applied the same γ estimation procedure as on real participant data. Results showed that our procedure sometimes underestimated and sometimes overestimated true γ, due to mapping of continuous perceived distance variable p_δ onto a discrete scale. See Supplemental Figure 8.4 for γ recovery on continuous p_δ values of the simulations.

3.3 Discussion

A large prior literature has used similarity ratings between exemplars as measures of inter-item distances in psychological representational space (reviewed in Chapter 1 sections 1.1 and 1.6), and multidimensional scaling (MDS) techniques have been used to depict low-dimensional visual representation of these relationships (e.g. Smith et al., 1974). Tversky and Gati (1982) pointed out that, often, such MDS reconstructions are taken as maps descriptive of the structural geometry of the underlying psychological representation, allowing inferences to be made on unobserved data. However, the accompanying requirements of such geometric models, such as the triangle inequality and segmental additivity, are often untested. Segmental additivity requires that along a straight line, segments are additive in a way that total distance is a sum of constituent distances. The triangle inequality requires that a direct path between two points be the shortest path; shorter than any indirect way going through a third point that does not lie on the path. Both are intuitive foundational requirements for any metric representational geometry.

In this chapter, we re-examined the classical geometric theories of knowledge representation for adherence to segmental additivity and the triangle inequality. Using ordinal scale measures of similarity, such as those generated in a pair-wise similarity judgment task, it is not possible to test for segmental additivity. However, the triangle inequality can be tested for ordinal data using the method developed by Tversky and Gati (1982). We employed this method to test pair-wise similarity judgment data for various 2D stimulus spaces, some of which we adapted from the recent neuroimaging literature, given that they have been claimed to support geometric models of conceptual representation (e.g. Balkenius & Gärdenfors, 2016; Bellmund et al., 2018). Additionally, we used one-dimensional optimization method to estimate the Minkowski γ parameter underlying the similarity judgments, giving us another way to check for the triangle inequality. Finally, we examined the validity of these ordinal and interval tests by running them on ideal observer similarity data, generated either from metric or non-metric spaces.

We designed six 2D stimulus spaces, grouped into three groups based on whether the stimuli were naturalistic or artificial and whether they were spanned by quantitative, qualitative, or a mix of quantitative and qualitative dimensions: (i) naturalistic bird stimuli with quantitative length dimensions, (ii) naturalistic plant and lamp stimuli with qualitative shape dimensions, and (iii) artificial stimuli defined by a qualitative and a quantitative dimensions. Although the data were inconclusive for Group 2 with lamps and plants, the bird stimuli of Group 1 satisfied ordinal triangle inequality, while the artificial stimuli of Group 3 violated it. We further found that our γ estimation procedure estimated the underlying Minkowski metric to be larger than 1 for all three of these groups.

3.3.1 Comparison with Tversky and Gati (1982)

These results conflict with those found by Tversky and Gati (1982) in several respects. These authors examined ordinal triangle inequality satisfaction and performed γ estimation for six different 2D stimulus spaces. Those 2D spaces that had psychologically separable dimensions were all found to violate ordinal triangle inequality and were associated with an estimated $\gamma < 1$. The 2D stimulus space with integral dimensions did not violate ordinal triangle inequality and had a γ estimate of more than 1. In our data, we expected Group 1 bird stimuli to be psychologically separable and thus to violate the triangle inequality, but we found they satisfied it. One possibility is that the quantitative bird space dimensions could have been perceived integrally, contrary to our intention (for example, the lengths of neck and legs could be combined nonlinearly into an overall

impression of size). The violation of ordinal triangle inequality by artificial stimulus spaces of Group 3 is consistent with the results of Tversky and Gati. However, our estimate of the Minkowski parameter for this group was not below 1. A closer inspection of the ordinal triangle inequality test procedure, coupled with our results with ideal observer simulations, sheds some light on this inconsistency.

3.3.1.1 Reasons for violating ordinal triangle inequality

The ordinal triangle inequality procedure (Equation 3.4) checks whether the two-dimensional distances are smaller than unidimensional ones in corner quadruplet stimuli (see Figure 3.1-B). If the physical-to-psychological distance mapping is linear (i.e. $\lambda = 1$; segmental additivity is satisfied), ordinal triangle inequality can be violated if the sum of the distances along the sides of the triangle is shorter than the distance along the diagonal. In other words, when combination of unidimensional distances into two-dimensional distance is superadditive. This is readily achieved with a power metric function with Minkowski γ parameter < 1.

However, ordinal triangle inequality can also be violated if $\gamma \geq 1$ but the physical-to-psychological mapping is not linear due to *unidimensional subadditivity* ($\lambda > 1$). For illustration, consider a corner triangle in Figure 3.1-A. If we have a city-block metric with $\gamma = 1$ but non-linear physical-to-psychological mapping with $\lambda > 1$, then $\delta(a,c) = \delta(a,d) + \delta(d,c)$, but $\delta(a,c) < \delta(a,b) + \delta(b,c)$. From this, it follows that $\delta(a,d) + \delta(d,c) < \delta(a,b) + \delta(b,c)$, i.e. the corner path is shorter than the centre path, leading to a violation of Equation 3.4. Thus, under conditions of inter-dimensional additivity or sub-additivity ($\gamma \geq 1$), the ordinal triangle inequality test developed by Tversky and Gati can reflect satisfaction or violation of segmental additivity and not of the triangle inequality. Indeed, a direct comparison of distances along the full diagonal $\delta(a,c)$ and the sum of the sides $\delta(a,d) + \delta(d,c)$ can satisfy additivity ($\gamma \geq 1$), while once the diagonal is broken down into constituent segments and those segments are individually compared to the sides of the corner triangle, we find violations of Equation 3.4 driven by unidimensional subadditivity ($\lambda > 1$). This offers one explanation for the differences between our results and that of Tversky and Gati: while in their data, ordinal triangle inequality was violated due to inter-dimensional superadditivity ($\gamma < 1$), given our estimates of $\gamma > 1$ our data point to intradimensional subadditivity as the likely reason ($\lambda > 1$).

3.3.1.2 Ideal observer simulations help clarify the inconsistency

These conclusions are further clarified by our simulations of ideal observer data characterized by various metric or non-metric distance functions. Application of ordinal triangle inequality tests to these data revealed that violation of ordinal triangle inequality (Equation 3.4) can be caused when $\gamma < 1$ and $\lambda = 1$ but also when $\gamma \geq 1$ but $\lambda > 1$. Thus, finding violations is incompatible with metric models (either $\gamma < 1$ or $\lambda > 1$, or both), while satisfaction is compatible with both metric ($\gamma \geq 1$, $\lambda = 1$) and non-metric models. Our estimate of γ for artificial stimulus spaces was above unity, suggesting that ordinal triangle inequality was violated due to unidimensional subadditivity with $\lambda > 1$. For the bird stimuli, because γ was estimated to be larger than 1 and ordinal triangle inequality was satisfied, we cannot exclude that they are metrically represented, although our data do not provide a definitive answer.

3.3.2 Role of attention in similarity judgment

As discussed in Chapter 1, researchers have proposed that attention could influence the choice of dimension along which stimuli are compared, even if this choice is not driven by external task requirements (Gärdenfors, 2000; Nosofsky, 1986, 1987, 1992b; Shepard, 1964; Smith & Heise, 1992). Notably, this might lead to similarity data violating the triangle inequality. Tversky and Gati (1982) discussed potential attentional confounds in their results, mentioning that there were no "apparent shifts" (p.150) in the frame of reference in the well-defined context of their experiments. Using mathematical derivations, they argued that random attentional fluctuations could not explain their data. However, they did not exclude the possibility of systematic attentional effects based on specific pair comparisons. Consider a triplet of stimuli {ap,cp,cr} in Figure 3.1-B. Assuming the underlying psychological space is actually metric with $\lambda = 1$ and $\gamma = 1$, then in the absence of any attentional biases, we should expect the triangle inequality to hold such that $\delta(ap,cr) = \delta(ap,cp) + \delta(cp,cr)$. However, it is possible that attention shrinks distances between those pairs that coincide along one of the dimensions. If this effect is strong enough, the sum of the triangle's sides will be shorter than the diagonal, violating the triangle inequality. Thus, the internal psychological representation of stimuli could be metric in nature, while additional attentional processes lead to similarity data that violate metric requirements.

Could ordinal triangle inequality violation of our artificial stimulus spaces of Group 3 be explained by such attentional processes? Our γ estimates for Group 3 stimuli indicated

satisfaction of inter-dimensional additivity ($\gamma = 1$), which would point towards the presence of violations of intra-dimensional additivity. Importantly, although attentional processes that selectively amplify certain dimensions could explain findings such as inter-dimensional superadditivity with $\gamma < 1$, they cannot explain phenomena within single dimensions such as intra-dimensional subadditivity with $\lambda > 1$. Furthermore, large prior literature indicates that, perceptual dimensions similar to those that spanned the 2D spaces of Group 3 stimuli have non-linear physical-to-psychological distance relationships. Such non-linearities have been documented for visual, auditory, tactile, and other types of judgment (Fechner, 1860; Houston & Shearer, 1930; Weber, 1851).

Thus, the likely explanation behind violations of ordinal triangle inequalities in our data stems from violations of intra-dimensional additivity. This suggestion, however, crucially relies on our γ estimates of > 1 being accurate for Group 3 data. Therefore, we validated it on ideal observer simulations testing whether estimated parameters correctly reflected the underlying generative process. We found that, in certain simulations, γ parameter was overestimated even when true γ was below unity. Importantly, this leaves open the possibility that our Group 3 stimuli were characterized by true $\gamma < 1$, which would invalidate classical geometric models but could be consistent with attention-weighed models. Therefore, while in combination with prior perceptual literature, the likely explanation for our data is non-linear physical-to-psychological mapping, our data cannot definitively exclude the applicability of augmented geometric theories such as the attention-weighted model.

3.3.3 Conclusions

To summarize, we present evidence that certain stimulus spaces violate axioms of geometric theories of psychological representation, and thus a different theory might be needed for an adequate algorithmic-level description of such representations. Although future studies are necessary to more definitively exclude possible role of attention, finding such violations of metric properties have consequences for studies (such as Theves et al. 2019) which use stimuli similar to our Group 3 spaces, and which find that brain regions (namely the hippocampus) represent abstract 2D spaces with an underlying Euclidean metric. Other stimuli, such as birds defined by quantitative dimensions and which have been used to argue for parallels between spatial and conceptual navigation (Constantinescu et al., 2016), could be represented geometrically, but the present data do not allow conclusive statements about this. Furthermore, we find that violations of the

ordinal triangle inequality method devised by Tversky and Gati (1982) could be indicative of violations of segmental additivity and not necessarily of the triangle inequality, in the likely presence of a non-linear physical-to-psychological mapping. Finally, our simulations illuminate the information loss that is accompanied by mapping of internal continuous psychological values onto a discrete ratings scale. Future studies using this method should aim towards using rating scales with sufficient range to allow participants to more accurately express internal perceptions.

4 Generalisation of Non-Spatial Schemas

4.1 Introduction

The previous 2 chapters examined validity of geometric theories of knowledge organisation which recently received indirect support from neural evidence finding parallels between spatial and non-spatial coding principles. In their review of this literature, Bellmund and colleagues (2018) outlined several outstanding questions in the field. One of them concerned the question of knowledge transfer: how does information acquired in one cognitive space facilitate (or inhibit) acquisition of related information in another one? Do the place and grid cells in hippocampal-entorhinal system that map one space get reactivated during knowledge transfer to a different space? Crucially, are these knowledge transfer dynamics similar when going between non-spatial domains versus when transferring between spatial and non-spatial areas? In this chapter, we attempted to design an efficient and flexible paradigm to capture such generalisation across different conceptual spaces, with the aim of extending it for studying transfer from conceptual to physical spaces or vice-versa. We conducted two experiments to validate our approach, but found that generalisation depended on ordering of our counterbalancing conditions, which were likely explained by characteristics of specific exemplar stimuli. We discuss these limitations and finish by outlining possible future lines of research that would help establish an efficient and flexible paradigm for testing knowledge transfer across both spatial and non-spatial domains.

4.1.1 Non-spatial schemas as structured representations of knowledge

Our knowledge about the world does not simply consist of isolated conceptual information. Objects, causal systems, landscapes, stories and other stimuli are not simply collections of features or qualities with certain values; these features and qualities often form propositional or hierarchical relationships with each other that need to be learned and captured in our internal world model (Markman, 2012). As discussed in the introductory Chapter 1, such rich structures have been characterised as *schemas*, or

networks of associative knowledge structures, which have been shown to have wide-ranging effects on various aspects of cognition such as memory acquisition, storage, and retrieval (Fernández & Morris, 2018; Ghosh & Gilboa, 2014; Gilboa & Marlatte, 2017; van Kesteren et al., 2012). What is more, such relational structures often repeat across situations. For example, hierarchies are common in social networks such as families or workplaces, while many processes in the world follow cyclical periodic structure such as the day-night cycle or seasonal transitions. Thus, schemas likely play a significant role in generalisation of knowledge between individual circumstances encountered by an agent (Taylor et al., 2021). Schemas could also be spatial, explicitly involving location information (Epstein et al., 2017; Farzanfar et al., 2022). In physical space, environments are often organized in repeated patterns, whether due to natural processes (such as plants growing next to water) or human-design consideration (such as designs of cities that often follow the same organisation).

The two algorithmic-level theories of knowledge representation discussed in the previous chapters – geometric theories and feature-based theories – are inadequate for learning and representing such relational knowledge. A proper representational format would need to explicitly encode relational structures, with variables that can take different arguments in different domains. Apart from the schema literature, related work encompasses research on development of semantic networks (A. M. Collins & Loftus, 1975; A. Collins & Quilliam, 1972; McClelland & Rumelhart, 1981), learning sets (Harlow, 1949), task models (Daw et al., 2005, 2011; Sutton & Barto, 1998), cognitive maps (Behrens et al., 2018; Peer et al., 2021), etc. Here, we focus on reviewing research on analogical reasoning (Holyoak, 2012), and how an efficient non-spatial schema generalisation paradigm could benefit our understanding of related psychological processes.

4.1.2 Generalization of knowledge during analogical reasoning

Generalisation of knowledge can take many forms (for an extensive review, see Taylor et al. 2021). The simplest form of stimulus-response generalisation is closely related to the concept of similarity that was used to study the structure of conceptual representations in the previous two chapters. In his seminal paper “Towards a Universal Law of Generalization for Psychological Science”, Shepard (1987) described how inter-item similarities within multiple stimulus domains (such as geometric shapes, consonant and vowel phonemes, Morse code signals, colours, etc.) closely predict the probability of generalizing a response across items. He demonstrated how response generalisation was

an invariant monotonic function of inter-item distances in similarity spaces (although much further work has uncovered a more complicated relationship between similarity and generalisation, e.g. Jones et al. 2006).

However, analogical knowledge transfer depends not only on superficial similarity comparison, but also on structural elements of the two knowledge domains that transcend their surface-level, perceptual features (Gentner, 1983; Gentner & Markman, 1997; Holyoak & Koh, 1987). Detection of such similarity is thought to crucially depend on explicit representation of role-based relations among the elements within a knowledge domain. As such, studies of analogical thinking fall under a larger topic of role-based relational reasoning, which in turn is closely related to broad aspects of human cognition such as inductive reasoning (Holland et al., 1986), causal inference (Cheng & Buehner, 2012; Holyoak & Cheng, 2011), problem solving (Bassok & Novick, 2012), etc. Thus, elucidating the neuropsychological underpinnings of analogical reasoning has wide-ranging implications for progress in understanding human cognition.

How do we retrieve relevant *source knowledge domain* from memory, align it structurally with the *target knowledge domain*, and perform inferences from prior knowledge to the new situation? Decades of studies (reviewed in Holyoak, 2012) have broadly delineated these processes, as schematized in Figure 1.3 in the introductory chapter. The "multiconstraint theory" developed by Holyoak and Thagard (1989) argued that multiple sources determine how the two domains align: (i) surface level perceptual or semantic similarity between elements, (ii) structural or role-based relational similarity, and (iii) pragmatic task-relevance and importance of functional roles of specific elements. The theory makes various predictions about how these constraints jointly guide analogical reasoning. When in conflict with each other, structural and functional alignment dominate over surface-level perceptual features. Surprisingly, contrary to the mapping process, the process of retrieving suitable analogs from long-term memory is governed by surface-level similarities (Gentner et al., 1993; Holyoak & Koh, 1987; Ross, 1989). This discrepancy has been an active topic of research, with various computational models developed to clarify whether mapping and alignment can be implemented as distinct or unified processes (Forbus et al., 1995; Hummel & Holyoak, 1997; Thagard et al., 1990). Finally, repeated analogical transfer can lead to development of an abstract schema devoid of information about specific sensory elements, which could independently operate during subsequent situations (Gick & Holyoak, 1983). However, abstract schema induction is not guaranteed and conditions that govern this are being actively investigated.

At the neural level, the *multiple-demand network* (MDN) has been identified as the key circuit supporting many of the processes required for analogical reasoning (Duncan, 2010). For example, in studies of fluid intelligence using Raven's Progressive Matrices Test, MDN is increasingly involved during problems that require integration of multiple relations (Christoff et al., 2001; Raven, 1938). Additionally, the most anterior part of the PFC called the frontopolar cortex has been noteworthy in its involvement during verbal analogy detection tasks of the form A:B and C:D (e.g. HAND:FINGER and FOOT:TOE). This region increases activity with semantic distance between the pairs, likely reflecting increased demands of relational reasoning (Green et al., 2006, 2010). Finally, the neighbouring orbitofrontal cortex (OFC) has been extensively implicated in representing task states, indicating its role in extraction and representation of task structure (Niv, 2019; Schuck et al., 2016; Wikenheiser & Schoenbaum, 2016; Wilson et al., 2014).

Interestingly, recent proposals have also discussed the potential role of hippocampal-entorhinal system in structure abstraction and representation. In line with the wider literature in analogical reasoning, Behrens and colleagues (2018) underlined the importance of *factorised representations*, i.e. decomposition of specific event representations into its content versus structure, and argued for the importance of explicitly representing structural relations in order to flexibly generalise across distinct situations. While overviewing the role of prefrontal regions, the authors also emphasized parallel functionalities in the hippocampal-entorhinal machinery that make it well suited for supporting factorisation. For example, fMRI activity in these regions tracks statistical transitions of discrete state-spaces (Garvert et al., 2017; Schapiro et al., 2013; Stachenfeld et al., 2017). Medial versus lateral regions of the entorhinal cortex seem to factorise the sensory input into its content and structure, respectively, which is later combined in a unified representation in HPC (Komorowski et al., 2009; Manns & Eichenbaum, 2006). Thus, during spatial navigation, grid cell activity in the medial EHC reflects the structure of the task, i.e. its two-dimensional Euclidean geometry. Analogously, in non-spatial tasks, such as concept learning and manipulation, grid cells might similarity extract the structure of the conceptual space (i.e. variation along relevant dimensions; e.g. in Constantinescu et al. 2016). Thus, Behrens and colleagues proposed that grid cells can track general patterns of abstract relations for any given task, and represent them as *basis sets* for describing this relational knowledge. A new problem can subsequently be captured by this basis set in order to allow novel inferences.

4.1.3 The current experiment

To summarize the currently open questions in the literature: (i) Analogical transfer of knowledge between distinct domains requires not only structural alignment, but also mapping of surface-level perceptual features (the multiconstraint theory); (ii) Analogical reasoning can sometimes (but not always) lead to abstract schema induction, though characterization of the precise conditions is an open field of enquiry; (iii) A crucial aspect of analogical reasoning is the extraction and representation of structure, supported by prefrontal regions and possibly the hippocampal-entorhinal circuit as well. Clarifying roles of these structures would also address the challenge posed by Bellmund et al. (2018) regarding transfer of knowledge between conceptual spaces that are supported by hippocampal-entorhinal activity.

To answer such questions, a fast and efficient generalisation task is needed that would allow systematic manipulation of surface versus structural similarities between domains, and examination of conditions that determine generalisation. We set out to develop such a paradigm based on geometric formulation of conceptual spaces. To this end, we adopted the 2D neck:legs space from Constantinescu et al. (2016), and asked participants to learn *paired-associates* (PAs) between specific bird-exemplars and target "reward" stimuli (pictures of Christmas objects). We then asked whether learning of similar PAs in a different bird space, defined by different dimensions, was facilitated when geometry of the PA arrangements was identical across the two bird spaces (*Congruent* group) versus when the arrangements differed (*Incongruent* group). Across two experiments, we found evidence consistent with generalization, but only for specific transfers from one arrangement of PAs to another. For Experiment 1, possible ceiling effects with learning some of the PAs might have obscured the benefits of congruency or costs of incongruency. While no such ceiling effects were apparent in Experiment 2, generalization still only occurred for one of the arrangement-to-arrangement conditions. We argue that further development of this paradigm will be fruitful for the study of generalization, and that a different stimulus set could be used to avoid any interactions between experimental factors.

4.2 Experiment 1

4.2.1 Methods

4.2.1.1 Participants

A total of 161 healthy young adult participants were recruited (90 females) from the prolific.co platform, aged 18-41 (M = 29.1, SD = 5.98), and paid £6/hour for their time, according to the Cambridge Psychology Research Ethics Committee protocol PRE.2020.018. Of these, 80 (43 females, 49.7% of those recruited) aged 18-41 (M = 29.15, SD = 6.01) passed the final quality and performance checks (see the section *Quality and performance checks* below) to be included in the data analysis.

4.2.1.2 Stimuli

The two conceptual spaces were the same as those used in Chapter 3 of this thesis: naturalistic birds varying in the length of neck and legs (adapted from Constantinescu et al., 2016) and another set of birds varying in the lengths of beak and tail (Figure 4.1 A and B). Pixel measurements of stimuli were the same as in experiment in Chapter 3.

The target stimuli were also reused from Constantinescu et al. (2016) and consisted of three pictures of toys: a Sledge (PA1), a Gingerbread Man (PA2), and a teddy Bear (PA3) (Figure 4.1-C). The same three targets were used for Phase 1 and Phase 2 of learning.

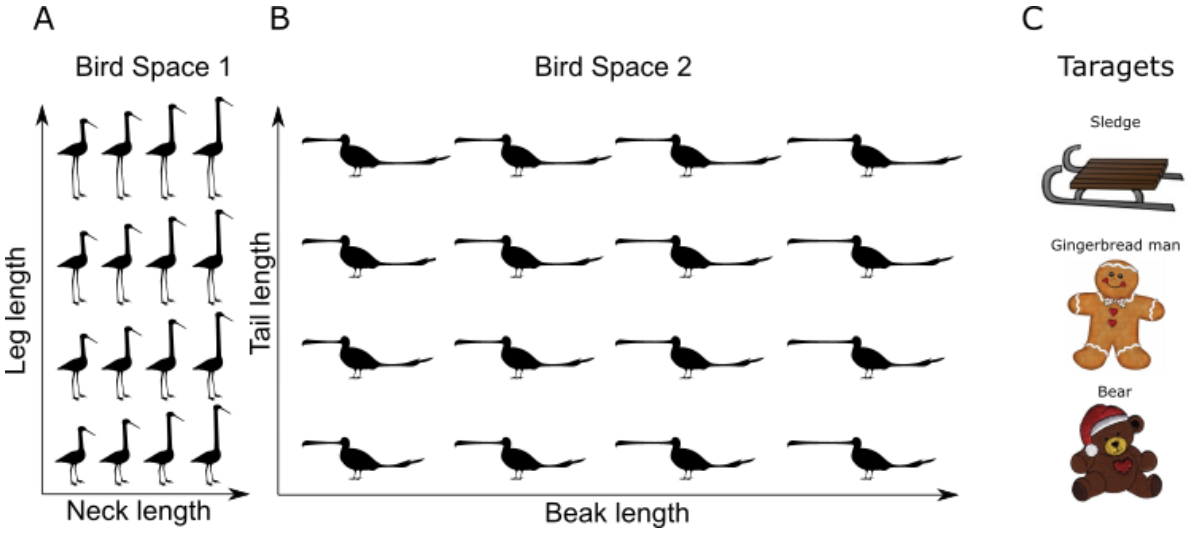


Figure 4.1: The two bird spaces and the targets.

(A) The first bird space, defined by a quantitative dimension of neck:leg length. (B) The second bird space, defined by a quantitative dimension of beak:tail length. (C) The target stimuli associated with specific bird exemplars, forming a paired-

associate (PA) that participants had to memorize. Stimuli in (A) and (C) reused from Constantinescu et al. (2016).

4.2.1.3 Arrangement of paired-associates

We created two arrangements of paired-associates (PAs) in the 2D space, shown in Figure 4.2. By "PA", we refer to the specific bird exemplar paired with a specific target stimulus. The PAs were at least two edges apart, did not share the same value for either of the two dimensions, and excluded the central locations to make them easier to learn. We attempted to choose *Arrangement 1* (*Arr1*) and *Arrangement 2* (*Arr2*) to be as different as possible from each other, to maximize incongruency across the learning phases. To this end, we designed Arr2 to avoid using the locations with the same coordinates in the second conceptual space, avoid having the same ordinal sequence of the target toys along either of the dimensions, and maximize the distance between each PA across the two conceptual spaces. However, due to experimenter error, the PA with Sledge (PA1) in Arr2 had coordinates that were mirror reflections across the 45 degree line of the PA with Gingerbread Man (PA2) in Arr1, resulting in lower incongruency between the arrangements than intended. This might matter since it is somewhat arbitrary how the axes in Arr1 map to the axes in Arr2, and so reflections across the 45 degree line could be conceived as the same location in both arrangements by some participants.

The mapping of target toys to locations in Arr1 and Arr2 was fixed across participants (e.g., the PA involving the Sledge in Arr1 was always paired with node 1, etc).

Crucially, none of the participants viewed this depiction of PAs distributed in the 2D conceptual spaces, but instead learned the bird-target associations through a learning procedure described below.

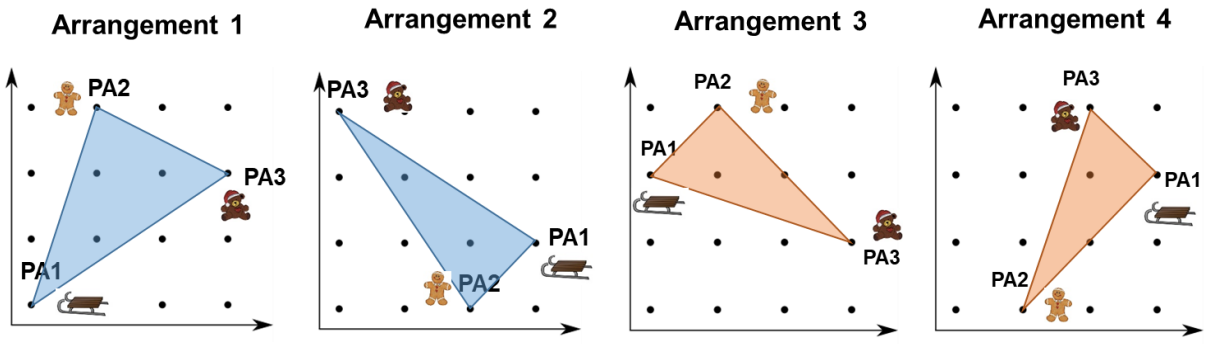


Figure 4.2: The arrangement of PAs used in experiments 1 and 2.
Left, the two arrangements of PAs used for Experiment 1. Right, the two arrangements of PAs used for Experiment 2. The participants never saw such a 2D

depiction of the stimulus space, but instead learned the PAs through rote trial-and-error learning (see the learning procedure below).

4.2.1.4 The learning procedure

4.2.1.4.1 Congruency manipulation

The experiment was broken up into two learning phases, one for each of the bird spaces. Concept order was counterbalanced across participants. Congruency of PA arrangements across the two learning phases was manipulated across participants: In the Congruent group, the participants started with either Arr1 or Arr2 (counterbalanced) in Phase 1 and continued with the same arrangement in Phase 2, whereas in the Incongruent group, participants had the arrangements switch across the two phases (Figure 4.3).

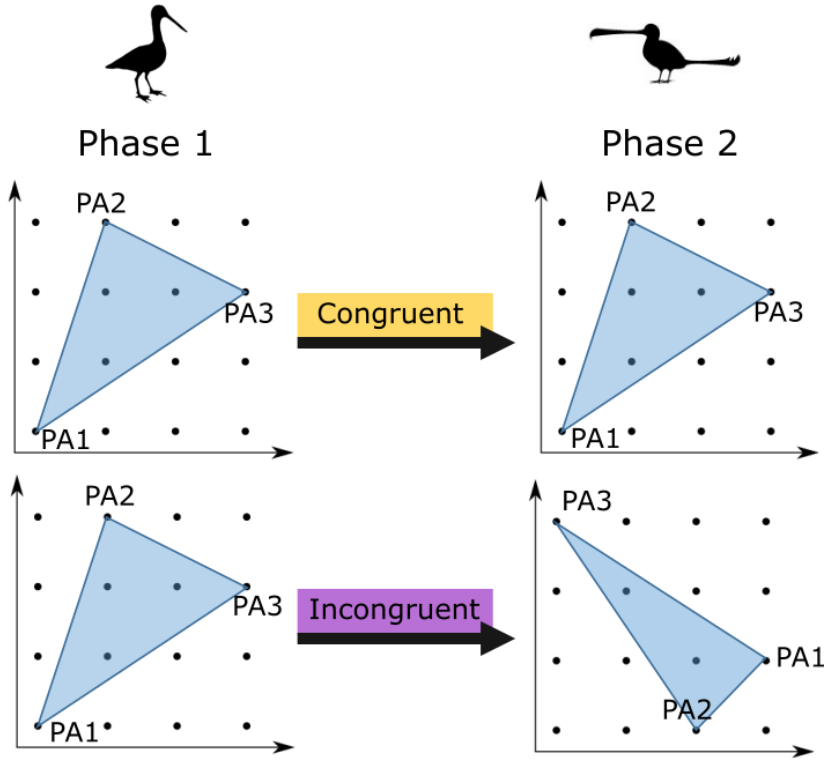


Figure 4.3: The main congruency manipulation.
For participants in the Congruent group, the arrangement of PAs was identical in the two bird spaces. For participants in the Incongruent group, Phase 2 had an arrangement of PAs different from that in Phase 1. Congruency was manipulated across participants, and the starting Arrangement and the bird space for Phase 1 was further counterbalanced across participants.

4.2.1.4.2 Task structure

The participants started by reading the task instructions and performing six practice trials for the first phase of learning. Next, they did Phase 1 of learning, followed by Phase 2 if they passed the quality and performance checks for the first phase. Each participant finished by completing a debriefing survey and being informed about the scientific goals of the study.

Each learning phase consisted of a minimum of two and a maximum of four learning blocks, with a small 15 seconds break between the blocks and a large 30 seconds break between the two phases. Each block consisted of 42 trials, with 14 trials for each of the three PAs. The 14 trials per PA included 13 trials with non-PA exemplars on the screen, and one trial with one of the other PA exemplars. On each trial, one of the three target stimuli were displayed on the screen along with two exemplar birds below it (Figure 4.4, Left). The two exemplars were offset horizontally to prevent exact comparison. The participants guessed which bird was the associate for the prompted target by pressing keyboard keys "1" or "2" corresponding to either the left or the right bird exemplar. The window of response was 30 seconds for the practice trials and 10 seconds for the learning trials. Feedback was given for 4 seconds by showing the targets associated with each of the two bird exemplars, and a "correct" or an "incorrect" statement with a green tick or a red cross for correct or incorrect responses respectively (Figure 4.4, Right). In case of missed response, the word "missed" was displayed for 4 seconds along with the correct bird PA. After 0.5 seconds from the feedback onset, the participants were allowed to press a space-bar to move onto the next trial. The ITI was set to 0.5 seconds. In the lower left corner of the screen, a scorebox showed the ongoing percent correct for each PA within a given block.

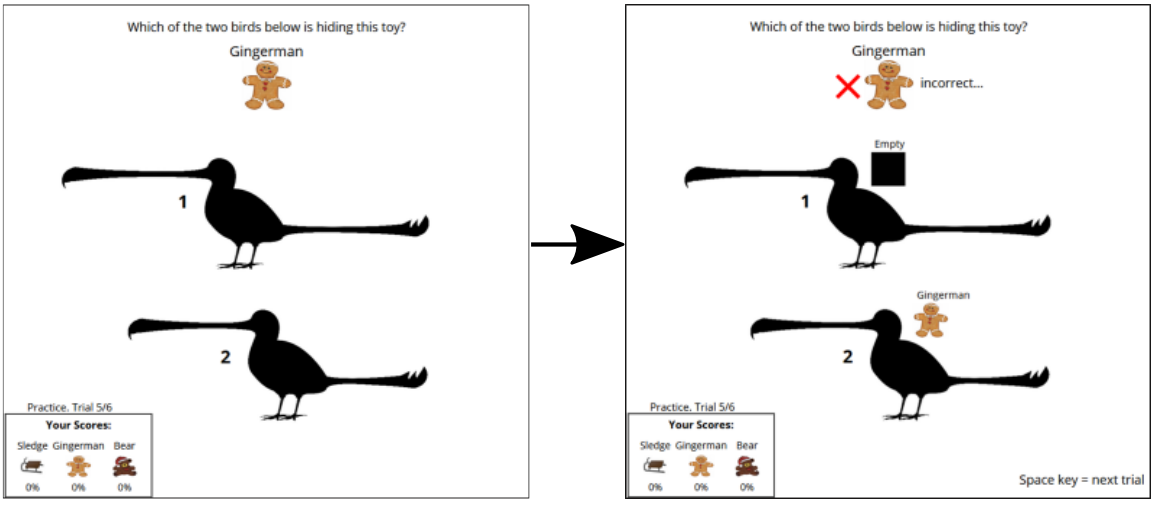


Figure 4.4: An example learning trial with feedback.
Participants were shown two bird exemplars on each trial along with one of the target stimuli as a prompt and asked to guess which of the two birds was the correct associate of the target stimulus. After a response, they received visual feedback on whether they were correct, along with the correct target associations for the two birds on that trial. A live tally of scores was always displayed on the screen.

4.2.1.5 Quality and performance checks

Quality and performance checks were conducted throughout the experiment. The participants had to reach 50% accuracy on each of the PAs by the third block and 85% by the fourth block in each of the two learning phases, otherwise they were discontinued from the study. The scores reset for every block within the learning phase.

Participants were discontinued if any of the following conditions occurred:

- Spent cumulatively less than 0.5 seconds on any of the instructions pages.
- Missed all of the practice trials.
- Responded faster than 1000ms for 85% of the trials for any block.
- Responded uniformly (either all "1"s or all "2"s) for 95% of the trials of any block.
- Missed more than 15% of the trials for any block.

4.2.1.6 Data analysis

We pre-registered the data analysis plan for this experiment on OSF: https://osf.io/w7f3g/

To capture learning in each phase, we measured two dependent variables: (1) the average performance in the first two blocks, and (2) the learning rate across all blocks. For the latter, we fit an inverse exponential function to the performance data of each participant, separately for each phase. The function was $y = 1 - intercept * e^{-c(t-1)}$, where y is the binary correct/incorrect values on each trial (missed trials were classified as incorrect), the intercept was set to 0.5 (chance) for both phases, c is the learning rate being estimated, and t is the trial index.

We calculated the acceleration of learning across phases by subtracting the Phase 1 values from Phase 2 values for each of the dependent variables.

As in Chapter 2, we used a Bayesian sequential design with maximal N procedure to assess evidence in favour of the alternative or the null hypotheses. H1 stated that the improvement in performance from Phase 1 to Phase 2 would be larger in the Congruent

compared to the Incongruent group, as assessed using a two-sided Bayesian t-test. H0 stated that there would be no difference between the congruency conditions. Maximum n was set to 136/group, such that the procedure would have a satisfactory power for supporting H1 and H0. Initial group size was 24/group, while batch size was 8/group. Criteria for BF_{10} and BF_{01} were set to 6.

We used Matlab R2020a (www.mathworks.com) and R RStudio (http://www.rstudio.com/) with R statistical software (R Core Team, 2022) for data preprocessing and analysis. Specifications for the parameters for the Bayesian t-test and the Bayesian ANOVAs can be found in section 1.7 of this thesis.

4.2.1.6.1 Power calculation

As in Chapter 2, we performed simulations to determine the power of our Bayesian sequential design procedure. Assuming a medium Cohen's effect size of d=0.5 (for comparing the amount of Phase 1-2 improvement across Congruent and Incongruent groups), with maximum n = 136 per group, our procedure had 93% chance of supporting H1 ($BF_{10} > 6$), 0.2% chance of supporting H0 ($BF_{01} > 6$), and 6.6% chance of remaining undecided. Assuming no effect (d = 0), our procedure had 64% chance to correctly support H0, 3.3% chance to incorrectly support H1, and 33% chance of remaining undecided.

4.2.2 Results

4.2.2.1 Quality checks and data exclusions

Our criterion of $BF_{10} > 6$ was reached after collecting 80 participants with valid data (40 Congruent and 40 Incongruent).

From the two dependent variables, the Phase 2 – Phase 1 differences in the exponential learning rate was highly skewed despite performing various transformations in the attempt to normalize it. Hence, all the subsequent analysis was performed on the second dependent variable, i.e. the Phase 2 – Phase 1 difference in the average performance in the first two blocks, which was roughly normally distributed.

4.2.2.2 The congruency effect

Figure 4.5-A shows the improvement in performance for each congruency group, i.e. Phase 1 – Phase 2 differences in average accuracy across the first two blocks. The BF_{10} for a t-test comparing the two groups was 269, indicating overwhelming evidence for H1 (Hedges g effect size assuming unequal variances g = 0.92). This suggests that the

participants in the Congruent group detected the shared structure between the two learning phases, which accelerated their learning in the first two blocks of Phase 2, i.e. showed generalisation. Figure 4.5-B shows the same data broken down in phases, showing a boost in performance in Phase 2 for the Congruent group (Hedges g = 0.91) while the Incongruent group scores did not improve (Hedges g = 0.02; i.e, little sign of generic practice effects on the task).

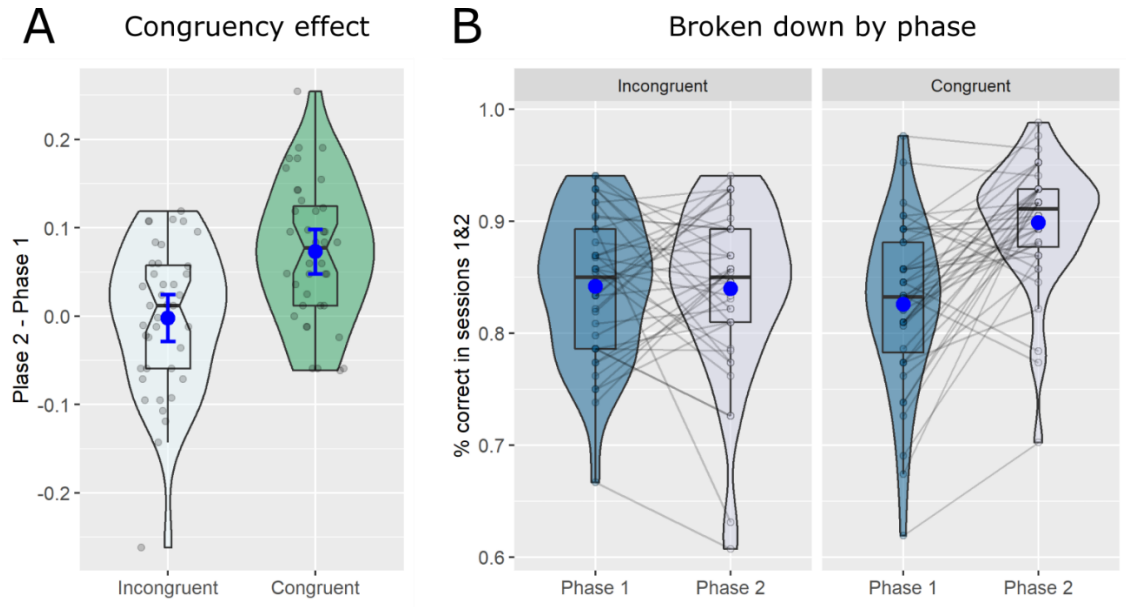


Figure 4.5: The congruency effect, Experiment 1.
(A) The congruency effect shown as a difference between the Congruent and the Incongruent groups in the amount of improvement from Phase 1 to Phase 2. Each dot is a participant. Blue dots represent group means. Error bars are 95% confidence intervals. (B) Performance broken down by each phase.

4.2.2.3 3-way Bayesian ANOVA to test for order effects and interactions

To test for any potential interaction between the observed congruency effect and any of the counterbalancing conditions, we conducted a three-way Bayesian ANOVA on the Phase 1 – Phase 2 difference scores for each participant. The three between-participant factors were: Congruency, Arrangement Order (Arr1 first versus Arr2 first) and Concept Order (Bird Space 1 first versus Bird Space 2 first).

Initial frequentist analysis of residuals was performed to check assumptions of the ANOVA. Shapiro-Wilk's test showed no evidence of non-normal distribution, and Levene's test showed no evidence of inhomogeneity of variances ($p > 0.05$ in both cases).

Bayes Factors were calculated for models that differed in their combination of main and/or interaction effects, relative to a null model with just a grand mean. A model with

all three main effects, plus a two-way interaction between Congruency and Arrangement Order, was strongly favoured over the null model ($BF=3.74x10^5$). It was 2 times more likely than the second-best model, which differed only in not including a main effect of Concept Order. This model was also 46.2 times more likely than a model including all three main effects but no Congruency:Arrangement Order interaction. This, combined with the observation that top 28 models (against the Null) all included a Congruency:Arrangement Order interactions, indicated that the effect of congruency depended on which arrangement the participants started with. This led us to follow-up with two separate 2-way Bayesian ANOVAs for each level of the Arrangement Order factor.

For the first level of the Arrangement factor (participants who started with Arr1), the winning model included the main effects of Congruency, with a $BF=1.24x10^5$ versus the null model. For the second level of the Arrangement factor, however, there was no evidence that any model was better than the null ($BFs<1.23$). Thus unlike participants who experienced Arr1 first, those who experienced Arr2 first did not show evidence of a Congruency or Concept order effect.

Plotting the results, Figure 4.6-A shows the congruency effect, separately for the participants that started with Arr1 and moved to either Arr1 (Congruent) or Arr2 (Incongruent) and for the participants that started with Arr2 and moved to Arr2 (Congruent) or Arr1 (Incongruent). The congruency effect is present for those who started with Arr1, but absent for those who started with Arr2.

To explore this further, Figure 4.6-B shows the same data, but now split by phases. For the Congruent participants moving from Arr1-to-Arr1 or Arr2-to-Arr2, we see the expected boost in performance in Phase 2. We also see an expected cost of incongruency for the participants going from Arr1-to-Arr2. However, for those going from Arr2-to-Arr1, we see an unexpected improvement in Phase 2, indicating that they found Arr1 in Phase 2 easy to learn despite having a different arrangement in Phase 1.

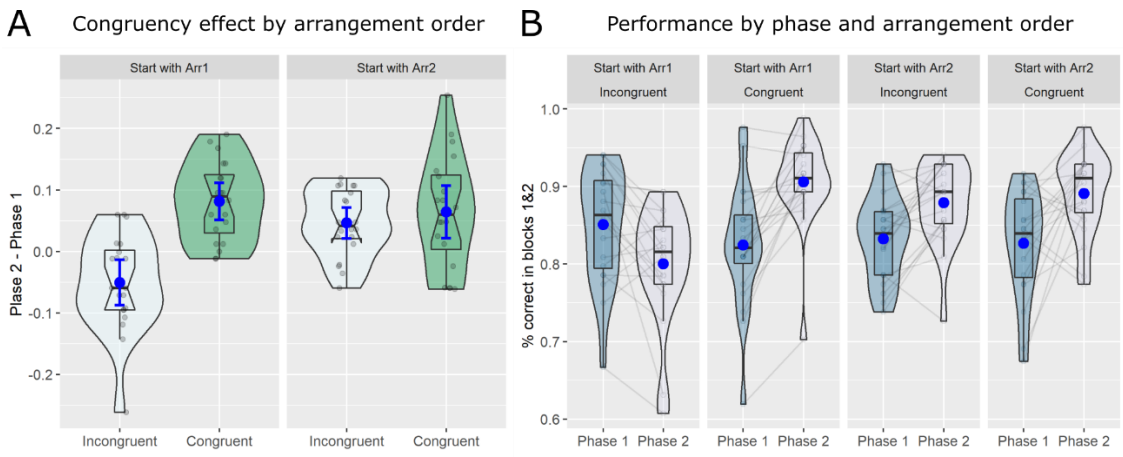


Figure 4.6: Interaction between Congruency and Arrangement order, Experiment 1.

(A) The Congruency effect broken down by Arrangement subgroups. The group starting with Arr1 showed a strong congruency effect, while the group starting with Arr2 did not. Each dot is a participant. Blue dots represent group means. Error bars are 95% confidence intervals. (B) Phase-by-phase breakdown of performance, split up by Congruency and Arrangement order.

4.2.2.4 The congruency effect is confounded by easiness of one of the PAs

To further understand this interaction between the congruency effect and the arrangement order, we looked at the scores for each of the three PAs separately. Figure 4.7 shows the mean scores across Blocks 1-2 for the Incongruent and the Congruent groups for each of the three PAs, broken down by Phase 1 and Phase 2 and by the arrangement experienced in that corresponding phase. We can see that for Arr1, the Sledge-PA1 was easiest to learn both in Phase 1 and in Phase 2. This might be explained by the fact that this exemplar bird had smallest features on all the dimensions (neck/legs or beak/tail, i.e, was in the "bottom left" of bird space), and was thus easily distinguishable from all the other exemplars, which might have aided in memorization. The participants in the Incongruent group who started with the Arr2 and moved to Arr1 seem to do exceptionally well on the Sledge-PA1 in Phase 2, which might boost their overall learning in Phase 2 and result in improvement of scores going from Phase 1 to Phase 2, despite being in the Incongruent condition. Thus, the ease of learning of one of the PAs in Arr1 might be masking the expected incongruency effect for the Arr2-to-Arr1 group. This is supported by the observation that effect size for congruency was larger when excluding the data for the outlier stimulus (Hedges g effect size assuming unequal variances g = 0.97, compared to 0.92 including the outlier). Our next experiment attempted to address this confound.

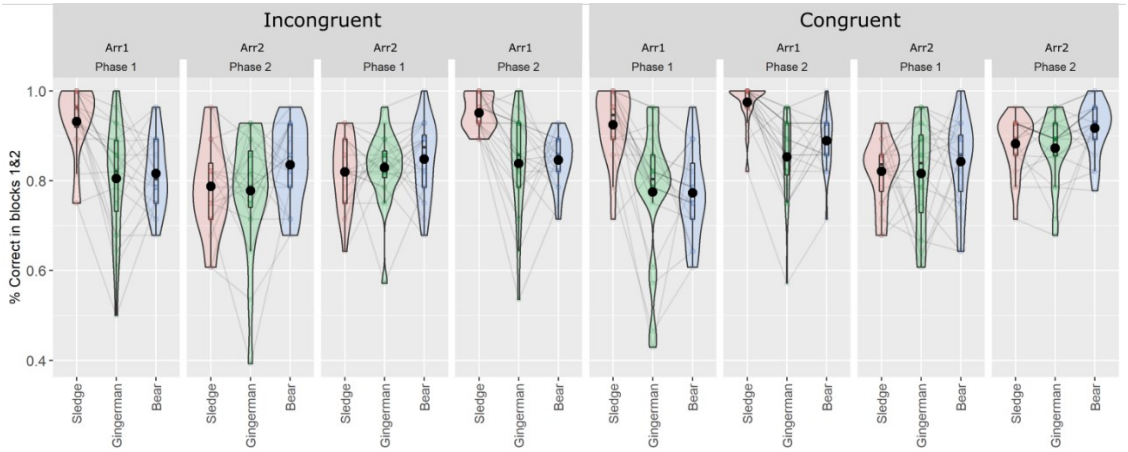


Figure 4.7: Performance by each PA, Experiment 1.

Average % correct in Blocks 1&2 for each of the 3 PAs, broken down by Congruency, Phase and Arrangement in the corresponding phase. The Sledge-PA1 stands out in Arr1, leading to close-to-ceiling performance whenever participants experience Arr1, regardless of the congruency condition.

4.3 Experiment 2

The second experiment involved the same setup as Experiment 1, except we generated a different pair of arrangements to use in Phase 1 and Phase 2, avoiding the lower left corner of the bird space (with smallest feature values), which Experiment 1 showed was especially easy to learn.

4.3.1 Methods

4.3.1.1 Participants

A total of 257 healthy young adult participants (162 females) were recruited from the prolific.co platform, aged 18-41 (M = 28, SD = 6.36), and paid £6/hour for their time, according to the ethics protocol PRE.2020.018. Of these, 126 (74 females) aged 18-41 (M = 28, SD = 6.3) passed the final quality and performance checks to be included in the data analysis (49%).

4.3.1.2 Stimuli

The stimuli were the same as for Experiment 1.

4.3.1.3 Arrangement of paired-associates

We created two new arrangements that satisfied the criteria set out for Experiment 1, but additionally avoided the lower-left exemplar bird with the shortest features. Figure 4.2

shows the two new Arr3 and Arr4 on the right, along with the two original arrangements on the left.

4.3.1.4 The learning procedure

Same as for Experiment 1.

4.3.1.5 Quality and performance checks

Same as for Experiment 1.

4.3.1.6 Data analysis

We used the same Bayesian sequential design with maximal n as for Experiment 1, calculating the BF_{10} and BF_{01} after each batch of participants for the same dependent variables as outlined above.

4.3.1.6.1 Power calculation

We performed the same simulation-based power analysis for a Bayesian sequential design with a maximal N as for Experiment 1. The expected effect size was set to d = 0.97, which corresponds to the effect size in Experiment 1 excluding the outlier stimulus. We found that with maximum n = 128 per group, 100% of the simulations resulted in support for H1 ($BF_{10} > 6$). In case of no effect (d = 0), 61.4% of our simulations supported H0, while 2.93% incorrectly supported H1, and 35.6% remained undecided.

4.3.2 Results

4.3.2.1 Quality checks and data exclusions

We reached the maximum number of participants without reaching either of the BF criteria. A total of 257 participants were tested, of which 49% (n = 126, with 64 Incongruent and 62 Congruent) passed both online quality plus performance checks and post-experiment debriefing checks. Of the 114 who did not pass online quality and performance checks, 11 additionally failed checks on their debriefing responses, while one participant experienced technical issues. Of the remaining participants who passed online checks, two experienced technical difficulties and 15 failed the post-experiment debriefing checks.

As for Experiment 1, all analysis was performed on the second dependent variable, the Phase 2 - Phase 1 difference in the average performance of the first two blocks.

4.3.2.2 The congruency effect

Figure 4.8-A shows the difference between the groups. Numerically, improvement in scores from Phase 1 to Phase 2 was larger for the Congruent group than the Incongruent group (Hedges g = 0.35). However, the Bayes Factor in favour of the alternative hypothesis failed to reach the criterion, $BF_{10} = 1.32$. Figure 4.8-B shows the data broken down into phases, demonstrating a slightly larger overall improvement from Phase 1 to Phase 2 for the Congruent (Hedges g = 0.53) than the Incongruent group (Hedges g = 0.11).

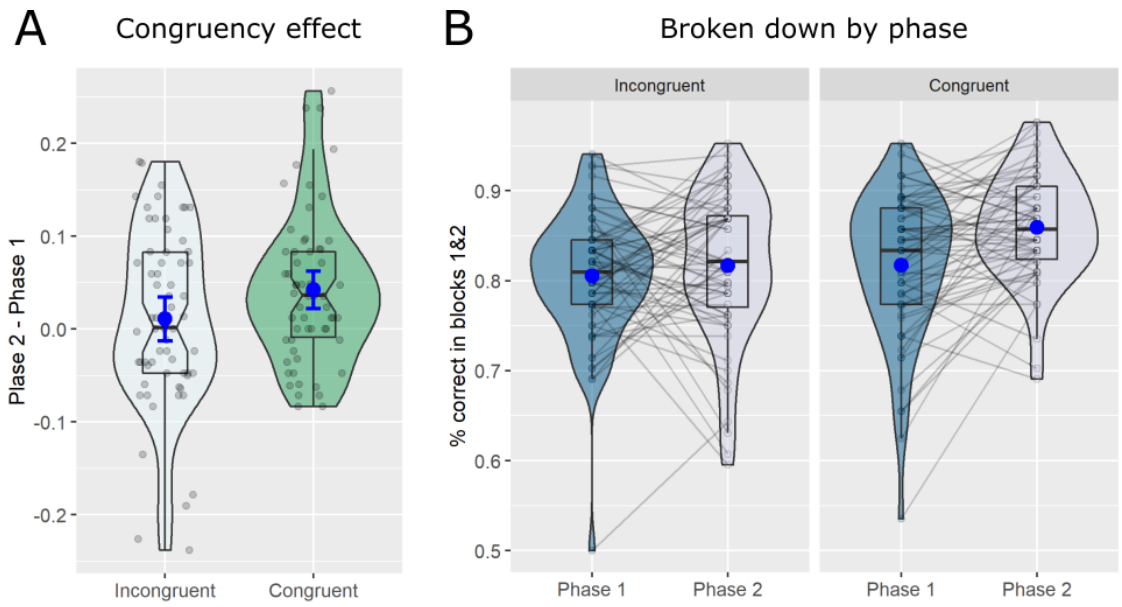


Figure 4.8: The congruency effect, Experiment 2.

(A) The congruency effect shown as a difference between the Congruent and the Incongruent groups in the amount of improvement from Phase 1 to Phase 2. Each dot is a participant. Blue dots represent group means. Error bars are 95% confidence intervals. (B) Performance broken down by each phase.

4.3.2.3 3-way Bayesian ANOVA to test for order effects and interactions

As for Experiment 1, we conducted a 3-way Bayesian ANOVA to test for any interactions between the congruency effect and other factors. Normality was assessed using Shapiro-Wilk's normality test and homogeneity of variances was assessed by Levene's test. Residuals were normally distributed ($p > 0.05$) and there was homogeneity of variances ($p > 0.05$).

The best ANOVA model included the main effects of Congruency, Concept order and the Congruency:Arrangement interaction, being >100 times more likely than the Null model

(BF_{10} = 101). This model was only 1.12 times more likely than the next best model, which only included the main effect of Concept order and Congruency:Arrangement interaction. Furthermore, this best model was 2.17 times more likely than the model including main effects of Concept order and Congruency without the Congruency:Arrangement interaction. Combined with the observation that all top 4 models (compared to Null) included the Congruency:Arrangement interaction, we decided to follow up with two separate 2-way Bayesian ANOVAs for each level of the Arrangement factor.

For the first level of the Arrangement factor, selecting participants starting with Arr3, the winning model included just the main effect of Concept order (Hedges g = 0.04) and was only BF_{10} = 1.05 times more likely than the Null. A model including the main effect of Congruency and Concept order had BF_{01} = 3.69, i.e. was *less* likely than the Null model, providing weak evidence for absence of Congruency effect for this subgroup of participants.

Next, we performed an analogous ANOVA on the second level of the Arrangement factor, selecting participants starting with Arr4. The winning model including main effects of Congruency (Hedges g = 0.69) and Concept order (Hedges g = 0.75) were BF_{10} = 77.1 times more likley than the Null model, and 2.77 times more likely than the next best model which additionally included a Congruency:Concept order interaction, and 7.5 times more likely than a model with just the main effect of Concept order providing strong evidence in favor for the main effect of Congruency.

Figure 4.9-A shows no difference between the Congruent and Incongruent groups for the participants starting with Arr3, while that difference was substantial for the participants starting with Arr4. Figure 4.9-B shows the breakdown by phases and reveals that, as in Experiment 1, the Congruent groups experience an improvement in scores from Phase 1 to Phase 2 as expected (Arr3-to-Arr3 Hedges g = 0.48; Arr4-to-Arr4 Hedges g = 0.55). For the Incongruent groups however, while Arr4-to-Arr3 shows the expected worsening in scores (Hedges g = 0.17), Arr3-to-Arr4 shows an unexpected improvement (Hedges g = 0.49).

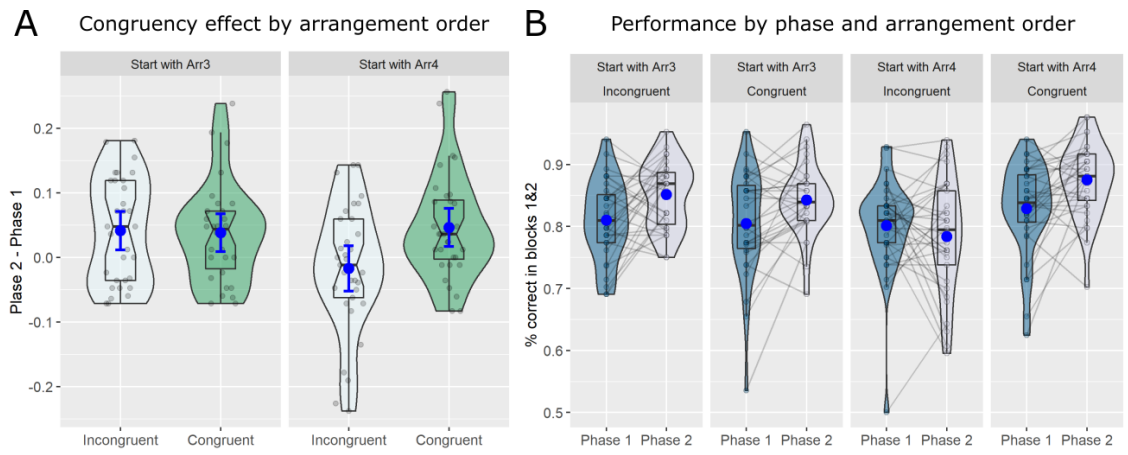


Figure 4.9: Interaction between Congruency and Arrangement order, Experiment 2.

(A) The Congruency effect broken down by Arrangement subgroups. The group starting with Arr4 showed a strong congruency effect, while the group starting with Arr3 did not. Each dot is a participant. Blue dots represent group means. Error bars are 95% confidence intervals. (B) Phase-by-phase breakdown of performance, split up by Congruency and Arrangement order.

4.3.2.4 No clear source for the interaction between congruency and arrangement order

We further explored whether this interaction was driven by any outlier PAs as in Experiment 1. Despite avoiding the lowest most left exemplar, Gingerbread Man-PA2 in Arr4 had small neck/legs that could have still been easier to learn compared to other PAs (see Figure 4.2 for Arr4 nodes). However, Figure 4.10, which shows the performance of the Incongruent group by each PA for Phase 1 and Phase 2, depending on the arrangement, did not suggest that any of the PAs were consistently easier or harder than the others. Thus, the source of the interaction between the congruency effect and the arrangement order in Experiment 2 remains unclear.

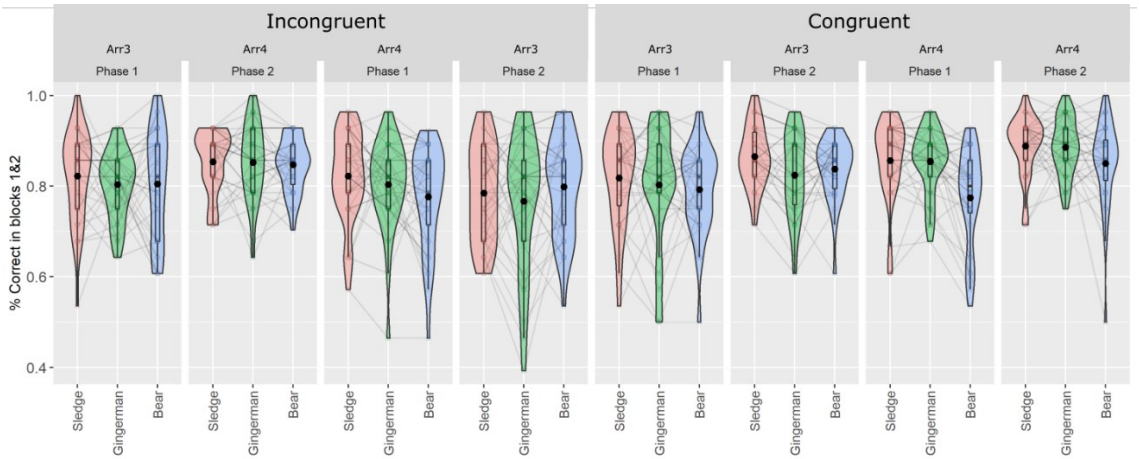


Figure 4.10: Performance by each PA, Experiment 2.
Average % correct in blocks 1&2 for each of the 3 PAs, broken down by Congruency, Phase, and Arrangement experienced in the corresponding phase. No clear outlier PA stands out that might have led to ceiling effects. Black dots represent within-group means.

4.4 Discussion

In this study, we aimed to develop a rapid, online generalization task involving analogical transfer of non-spatial schematic knowledge across two conceptual spaces defined by different quantitative dimensions. Such a paradigm would allow systematic examination of factors influencing successful knowledge transfer, with a possibility to adapt it to examine conceptual-to-physical knowledge transfer and contribute to elucidating recently proposed parallels between spatial and non-spatial processing (e.g. Bellmund et al., 2018; Morton & Preston, 2021; Moser et al., 2017). Across two experiments using a pair of 2-dimensional bird spaces based on a previous study (Constantinescu et al., 2016), we found evidence consistent with generalization, but only for some of our counterbalancing subgroups involving specific transitions of arrangements from Phase1-Phase2.

In Experiment 1, we found that the Congruent groups experienced an expected boost in Phase 2 regardless of their Phase 1 starting arrangement (Figure 4.6). Furthermore, one Incongruent sub-group showed an expected decline in Phase 2 performance. However, the other Incongruent sub-group showed an unexpected improvement in Phase 2 performance.

A possible explanation for the latter surprising result concerns the effect of an outlier stimulus on learning. The paired-associate involving the bird-exemplar with the smallest features might have allowed for easy discrimination from other birds, explaining why performance on this "corner" outlier stimulus was higher than that of the others (Figure 4.7). Thus, the Incongruent sub-group that transitioned to the arrangement with this outlier exemplar still performed exceptionally well overall, potentially obscuring any incongruency costs. Under this interpretation, the Congruent groups experienced genuine benefit of generalizing their knowledge of the arrangements from Phase 1 to Phase 2.

An alternative interpretation is that there were no congruency effects in Experiment 1, with the general improvement in Phase 2 for the Congruent group being attributed to simple practice effects (Harlow, 1949). Such generic practice effects would also be expected for the Incongruent group, in which case the unexpected result was instead those

in the Incongruent sub-group who showed worse performance in Phase 2. It is possible that, because this sub-group learned the arrangement with the outlier "corner" exemplar in Phase 1, they had less room to show any improvement in Phase 2.

In Experiment 2, we attempted to remove this confound by excluding any extreme "corner" exemplars when choosing arrangements. However, as in Experiment 1, we still observed an Arrangement-by-Congruency interaction. One Incongruent sub-group showed a decline in Phase 2, consistent with some generalization of (conflicting) information, but the other showed an unexpected improvement in Phase 2. A closer examination of the new arrangements chosen for Experiment 2 (Figure 4.2) revealed that similar to Experiment 1, there was one potential PA near the "edge" of the bird space, for which ceiling-level performance could obscure any incongruency costs. However, accuracy on this PA was not markedly higher than on other PAs (Figure 4.10). Thus, the source of the observed arrangement-by-congruency interaction in Experiment 2 remains unclear.

Given these results, we propose that further optimization of stimulus sets is necessary to continue establishing a fast and reliable behavioral paradigm for the study of generalisation of structured knowledge across conceptual domains. For example, a wider range of variation could be implemented along the two feature dimensions defining the conceptual spaces, which might allow avoidance of exemplars with ceiling or floor effects. Establishing such a working paradigm for the study of generalisation opens door to a plethora of subsequent hypothesis examinations and interesting manipulations, some of which we outline below.

First, generalisation could be examined in the context of varying the type of dimension defining the conceptual spaces. As introduced in Chapter 1, dimensions can be quantitative (length, size) versus qualitative (shape, colour), or psychologically integral (e.g., hue and chroma) versus separable (e.g., size and orientation) (Garner, 1976; Gati & Tversky, 1982). Systematically characterizing the influence of such factors on generalisation of knowledge will contribute to a fuller understanding of the underlying psychological processes.

Second, a large space of fruitful experimental manipulations is possible with regard to the structural similarity between the two conceptual spaces, through variation in the type of arrangements. Instead of having two extremes, one could gradually vary the level of congruency between the learning phases, such as rotating the arrangements by 90^0, or

laterally shifting them in space while maintaining their geometric shape. Such variations in structural similarity can be combined with variations in surface level perceptual similarity between domains, which can systematically test predictions of analogical theories such as the multiconstraint theory of Holyoak and Thagart (1989) discussed in the introduction. Additionally, certain arrangements might encourage adoption of verbalizable *rule-based* strategies for learning PAs, requiring attention to only one of the dimensions to achieve reasonable performance (Ashby et al., 1998). Other arrangements might require integrating across both dimensions, leading to an adoption of implicit *information-integration* strategies. Previous literature has found that humans can generalize a categorization rule in tasks requiring verbalizable rules, but not implicit information-integration strategies (Casale et al., 2012).

Third, one could compare different ways of teaching the PA arrangements to the participants. Specifically, one might expect different representations to support a 2D space depending on whether learning is via the rote trial-and-error approach used here, versus the navigation-like bird-morphing procedure used by Constantinescu and colleagues (2016). In that study, the participants used dials to smoothly morph birds into each other while setting a specific neck:legs morph ratio, thereby "navigating" in the 2D bird space at a certain angle and "discovering" various target toy stimuli that were associated with specific bird exemplars. It is possible that such different learning strategies will be conducive to different levels of generalization across conceptual spaces.

Fourth, one could examine whether the knowledge of non-spatial schemas consisting of arrangement of stimulus paired-associations in a conceptual 2D space would generalise to a task involving a spatial schema consisting of arrangement of actual landmark PAs in a physical environment, and vice-versa. Such a paradigm would accelerate the study of boundary conditions for links between spatial and non-spatial knowledge domains and their underlying neuro-computational mechanisms. Examination of effects of spatial schematic knowledge on learning is taken up in Chapter 5.

Finally, an efficient paradigm to capture knowledge transfer across tasks would be helpful for study of the neural basis of such generalisation (Taylor et al., 2021). As discussed in the introduction, specific contributions of prefrontal regions as well as those of the hippocampal-entorhinal system could be characterised, clarifying the neural processes underlying structural abstraction and inference.

In summary, given its central importance in human cognition, generalization needs to be systematically and carefully studied in a controlled experimental paradigm. We believe it is important to develop a fast and efficient learning paradigm to allow such investigation, and hope that subsequent research can succeed in optimising the stimuli and experimental conditions fit for this purpose.

5 Spatial schemas and Their Influence on Learning

5.1 Introduction

As discussed in the previous chapter, repeated generalisation can lead to induction of an abstract schema that can further facilitate knowledge transfer across domains. At the same time, much of psychological research on schemas has focused on how they facilitate integration of new within-domain knowledge (e.g. Tse et al., 2007, 2011; van Buuren et al., 2014; van Kesteren et al., 2010, 2013; Wang et al., 2012; for reviews see van Kesteren et al., 2012; Fernández & Morris, 2018; Ghosh & Gilboa, 2014; Gilboa & Marlatte, 2017). In this final empirical chapter of my thesis, we present a brief overview of relevant studies that show how spatial schemas influence learning, and present two experiments that suggest that extant data might be explained in other ways, such as the encoding of the location of single landmarks, independent of other objects in the space.

We define a schema as an interconnected network of associative knowledge structures, which in the case of a spatial schema, consists of typical locations of objects in relation to other objects (e.g., that a taxi company is often found near a train station in a city). Examination of precise neurobiological basis of such processes was first initiated by a rodent study of Tse and colleagues (2007). These authors taught rats flavour-place paired-associations (PAs) in a 2D arena, such that presence of a certain flavour predicted location of food in a particular sand-well (see Figure 5.1-A). In the *consistent* condition, the flavour-place associations (i.e. the PAs) remained stable across training days (Figure 5.1-B). In the *inconsistent* condition, the same rats were trained on another set of PAs in a different 2D area in a different room. More specifically, the flavour-place associations would swap every third training session. Although food locations remained stable, swapping of flavour-place pairings rendered the schema inconsistent. During a crucial test day, the rats learned a new flavour-place association in each room, with the new PA being adjacent to one of the old food locations (Figure 5.1-B, panel “New-PA Learning”).

The authors showed that, compared to the inconsistent condition, learning of the new PA was much faster in the consistent condition. This result has been interpreted as facilitatory influence of schemas as a network of associative knowledge.

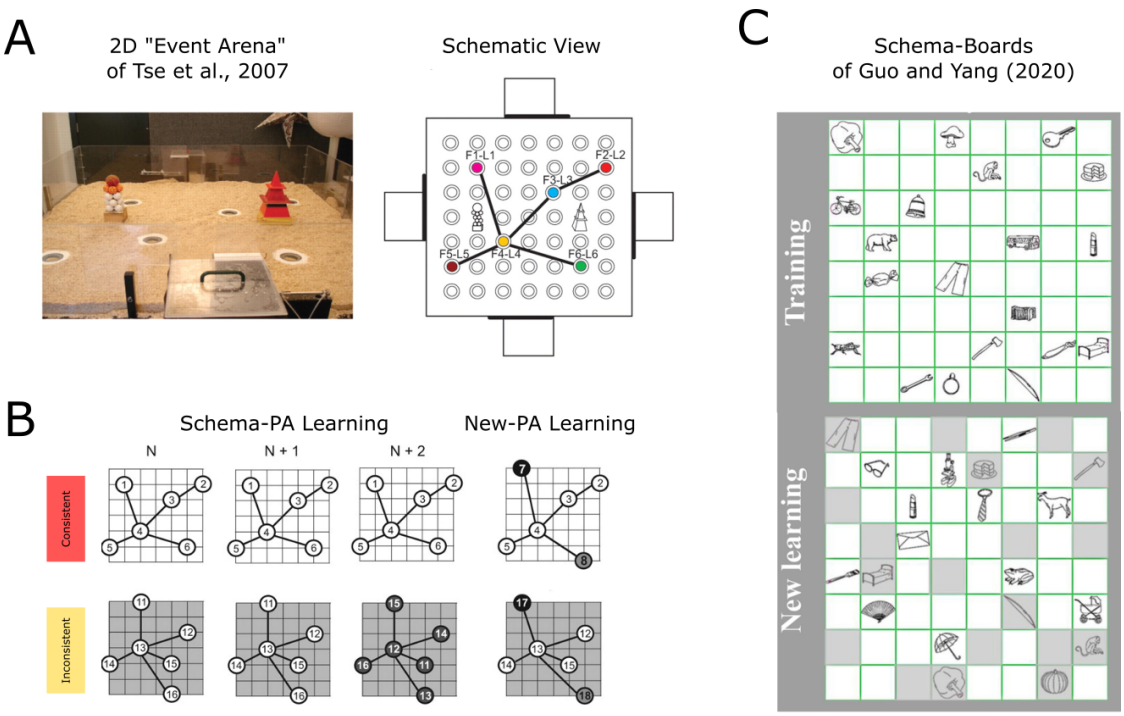


Figure 5.1: Schema paradigms of Tse et al. (2007) and Guo and Yang (2020). (A) The schema task of Tse et al. (2007). Left: rats learned flavour-place associations in a 2D event arena comprising of sand-wells where food might be hidden and proximal and distal cues for orientation. Right: A schematic depiction of the event arena and associations between flavours (F1, F2, etc) and locations (L1, L2, etc). (B) The Schema-Learning and the New-PA Learning stages for the Consistent and Inconsistent schema groups. Note that at the New-PA Learning stage, locations of the New-PA sand-wells are adjacent to locations of previously learned PAs. (C) Example boards used by Guo and Yang (2020). PAs consisted of image-location associations on 8x8 boards. In the Schema-Consistent condition, these associations remained stable, while in the Schema-Inconsistent condition, the images swapped locations at the beginning of each training day. At the New Learning stage, participants learned 12 new-PAs together with 8 old-PAs. Grey squares denote locations of old-PAs (not greyed out during the actual experiment). Panels (A) and (B) adapted from Tse et al. (2007). Reprinted with permission from AAAS. Panel (C) from Guo and Yang (2020). Reprinted with permission from John Wiley and Sons.

One unanswered question is whether in the Tse et al. (2007) paradigm, the schema-defining PAs acted as a unified global network to influence learning, or whether each PA was encoded independently and locally facilitated learning of a new PA that happened to be nearby. Indeed, the new PAs that the rats learned in each room were directly adjacent to old PA locations, making it impossible to disentangle such "global" versus "local" influence of schema elements.

The same shortcoming applies to studies that have adapted the paired-associates learning task for humans. In one such recent adaptation, Guo and Yang (2020) taught their participants image-location associations on a 2D grid board on a computer screen (Figure 5.1-C). In the consistent schema condition, the image-location PAs remained stable across training days, whereas in the inconsistent condition, the locations where images would appear remained the same, but image-location mappings would shuffle at the start of each training day. The authors found that, during a subsequent new PA learning stage, the new PAs were learned better on the consistent boards than on the inconsistent ones, demonstrating the facilitatory effect of schemas on new learning. As with Tse et al. (2007), every new PA in Guo and Yang's paradigm was also immediately neighbouring an old PA location, making it impossible to disentangle a local facilitatory effect of independently encoded old PAs, or the global influence of the interconnected network of old PAs.

We designed an experiment where we directly manipulated whether the to-be-learned PAs were next to a neighbouring schema item or far from it. Participants learned locations of hidden images (the *Hidden-PAs*) on boards consisting of 12x12 grids (Figure 5.2). Five within-participant learning conditions allowed us to disentangle global versus local influences as well as test for various other hypotheses as outlined below.

5.1.1 The local versus global influence of associative elements

In every learning condition, participants had to find 6 Hidden-PAs on a board through trial and error. In the consistent schema (*Schema-C*) condition, the board also contained 6 *Visible-PAs*, which were displayed on the board at the beginning of each trial (Figure 5.2-A). These Visible-PAs remained in consistent locations throughout training, being directly available for the participants as landmarks to scaffold learning of the Hidden-PA locations. Thus, unlike previous studies, our paradigm had no separate schema-learning stage, instead having the Visible-PAs directly presented as a stable knowledge structure, i.e. as a schema. To test the "local versus global" hypothesis, some of the Hidden-PAs

(called *Near-PAs*), were located directly adjacent to some of the Visible-PAs, while others (*Far-PAs*) did not have any close-by Visible-PAs (Figure 5.2-A). If Hidden-PA learning is facilitated by local influences of such individual "landmarks" (i.e. individual Visible-PAs), then Near-PAs should be learned better than Far-PAs. If, on the other hand, Hidden-PA learning is influenced globally by the network of all Visible-PAs (i.e. a spatial schema), then Near- versus Far-PA performance should be equal. This Near versus Far-PA comparison within the Schema-C condition formed the first of the two primary contrast for our experiment.

Even if Near-PAs are learned better than Far-PAs in the Schema-C condition due to local influence of nearby Visible-PAs, it is possible that far-away Visible-PAs still exert some beneficial effect from a distance. To disentangle this effect, in a third condition – the *Schema-Landmark* (*Schema-L*) condition – two of the six Visible-PAs remained stable (i.e. acted as landmarks) whereas the remaining four moved randomly anywhere on the board (Figure 5.2-A). As with the Schema-C condition, the Schema-L condition had two Hidden-PAs that were Near-PAs, being adjacent to the two stable, landmark Visible-PAs. If the local influence hypothesis is true, and only the nearby Visible-PAs facilitate learning, the Near-PAs in Schema-L condition should be learned as well as Near-PAs in Schema-C. If, on the other hand, the far-away Visible-PAs in Schema-C that remain stable across trials have some beneficial effect at a distance, the Schema-C Near-PAs will be learned better than the Schema-L Near-PAs. This comparisons of Near-PAs in the Schema-C and Schema-L conditions formed our second primary contrast for this study.

5.1.2 The location knowledge hypothesis

Apart from the two main contrasts outlined above, we examined several secondary hypotheses. Similar to previous schema studies, we included a *Schema-Inconsistent* (*Schema-IC*) condition where the slots for Visible-PAs remained stable across trials but all the images swapped places with each other on every trial (Figure 5.2-A). Despite the swapping, participants could develop a *location knowledge*, where they remember the location of the objects, abstracted away from the objects themselves, and use this to scaffold learning of Hidden-PAs. This location knowledge might be just as effective as the *image-location knowledge* in the Schema-C condition. We thus compared overall learning in Schema-C and Schema-IC conditions. We also compared Near versus Far-PA learning within the Schema-IC condition, to see whether, even if schemas are locations only, knowledge of the spatial relationships between those locations help both Near and

Far-PA learning, or does new learning benefit just from single landmarks (encoded by their relationship to the borders of the grid), which are independent of other locations (so not really a schema), and which only help Near-PAs.

5.1.3 The distraction hypothesis

Given that in our paradigm, the Visible-PAs appear at the beginning of every trial, it is possible that random relocation of PAs across trials might lead to distraction and hinder learning. To test for such a *distraction hypothesis*, we included two final learning conditions (Figure 5.2-A). In the *Random* condition, all the Visible-PAs moved randomly anywhere on the board on each trial. In the *No-Schema* condition, no Visible-PAs were present, and participants only learned locations of the Hidden-PAs. If randomly moving Visible-PAs on every trial have a distracting effect, the Random condition should do worse than the No-Schema condition. Note that comparison of Random and No-Schema conditions was not part of our pre-registered secondary analysis, but is particularly suited for testing the distraction hypothesis[4].

Panels B-D of Figure 5.2 present the predictions for the three hypotheses outlined above (across the 5 conditions in Figure 5.2-A, and the Near versus Far comparison within each condition). If the "local influence" hypothesis is correct, we expected that Near-PAs would be learned better than Far-PAs within the Schema-C condition, and Near-PAs of Schema-C would be equal to Near-PAs of Schema-L. If the "location knowledge" hypothesis is true, performance in Schema-IC will be comparable to Schema-C, and Schema-IC will also show a Near-Far advantage for the Hidden-PAs. Finally, if the "distraction hypothesis" is correct, we would expect that the Random condition would do worse than the No-Schema condition. Of course, it is possible that more than one factor is at play (i.e. more than one hypothesis is true), in which case the data might be a combination of several such predictions.

[4] The pre-registration originally planned to compare Random with Schema-L condition. However, because these two conditions differ in both, the number of randomly moving Visible-PAs *and* stable Visible-PAs any difference in performance could be due to either distraction or global effect of extra stable landmarks.

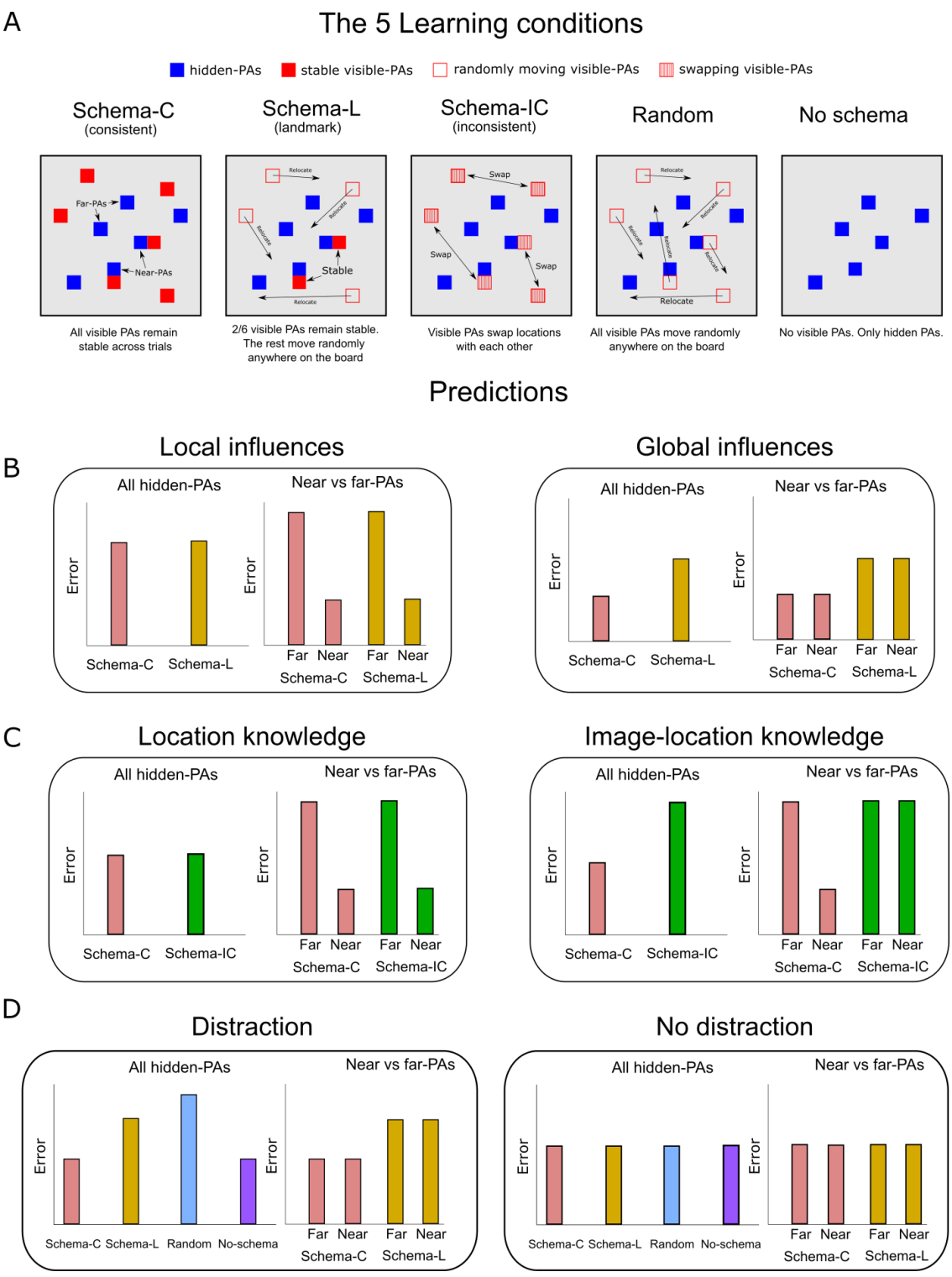


Figure 5.2: Experiment 1 Learning conditions and predictions.

(A) There were 5 within-participant conditions, where participants had to find 6 Hidden-PAs on a board. The conditions differed based on whether the Visible-PAs remained stable, swapped places, or moved randomly anywhere on the board across trials. (B) Predictions for the "local vs global" hypotheses. Under the "local influence" hypothesis, we expected no overall difference between Schema-C and Schema-L and a difference between the Near-PAs and Far-PAs within the Schema-

C condition. Under the "global influence" hypothesis, Schema-C would outperform Schema-L overall, while there would be no difference between Near-PAs and Far-PAs. (C) Predictions for the "location knowledge" hypothesis. If knowledge of stable locations in the Schema-IC condition is enough to benefit learning, we expected equal overall performance in Schema-C and Schema-IC, and a Near-PA versus Far-PA difference for both conditions. If image-location stability is necessary, however, we expected Schema-C to outperform Schema-IC, while Near-PAs would outperform Far-PAs in Schema-C only. (D) Predictions for the "distraction" hypothesis. Distraction would not impair learning in the Schema-C and the No-Schema conditions, while effecting the Random condition worse than the Schema-L condition. If distraction was not a factor, we expected no difference between the overall learning in any of the conditions. Regardless of the effects of distraction, we predicted no difference between the Near-PAs and the Far-PAs in Schema-C or Schema-L. Note that real data might, of course, be a combination of several such predictions, if more than one such factor is at play.

5.2 Experiment 1

5.2.1 Methods

5.2.1.1 Participants

86 healthy young adult participants were recruited (41 females) from the prolific.co platform, aged 18-40 (M = 30.14, SD = 6.16), and paid £6/hour for their time, according to the Cambridge Psychology Research Ethics Committee protocol PRE.2020.018. Of these, 65 (34 females, 30 males) aged 19-40 (M = 30.39, SD = 6.22) passed the final quality and performance checks (see the *Quality checks* Section below) to be included in the data analysis. One participant that passed the QC checks withdrew their age and gender information from prolific.co.

5.2.1.2 Stimuli

We used a 2D board consisting of a grid of 12x12 locations in which images could be placed to form image-location paired-associates (PAs). Unlike the previous human schema paradigms (Guo & Yang, 2020, 2022; van Buuren et al., 2014), no grid lines were shown, in order to minimize the use of explicit verbal strategies in encoding PAs based on rows and columns.

For the PAs, images were chosen from a bank of standardized stimuli (Brodeur et al., 2014), and consisted of everyday household items, animals, plants and various man-made objects. The images were filtered to have high familiarity (with the parameter "Familiarity Mean" > 4) according to the mean familiarity ratings provided by Brodeur et al.

5.2.1.3 Task design and procedure

There were five within-participant learning conditions, each containing a square board (Figure 5.2-A). The colour of the board was different for each learning condition. Each board had 6 Hidden-PAs that the participants had to find through trial-and-error. The arrangement of Hidden-PAs differed across the boards, and the arrangement-to-condition assignment was counterbalanced across participants (see below).

4 of the 5 learning conditions additionally involved presence of 6 Visible-PAs on the boards, which acted as landmarks or a schema that could be used for learning Hidden-PA locations. The 5th condition had no Visible-PAs.

5.2.1.3.1 Trial structure

Instead of having a separate schema-learning stage as in previous human schema tasks (Guo & Yang, 2020, 2022; van Buuren et al., 2014), the participants directly saw the 6 Visible-PAs at the beginning of each trial (Figure 5.3-C). On each trial, the Visible-PAs were displayed on the board for 2 seconds (for the 5th condition, only an empty board was shown). Then, the board disappeared, and a prompt image appeared for 500ms indicating which of the 6 Hidden-PAs to find on that trial. Following this, an empty board reappeared without Visible-PAs and the participants indicated the location of the prompted Hidden-PA with a mouse click. The response window was 3 seconds. Following a response (or after maximum time had elapsed), the Hidden-PA of that trial and all six Visible-PAs appeared on the board, along with feedback on accuracy (correct versus incorrect versus missed trial). Following an ITI of 500ms, the next trial began.

On each trial, the location of the board itself was randomly varied within a central area, to avoid the participants using dirt marks on their computer screens as alternative landmarks.

Each Hidden-PA trial was repeated eight times, resulting in 48 trials for each condition broken up over two learning blocks (Figure 5.3 panels A and B). A small break was given between the blocks and between the learning conditions.

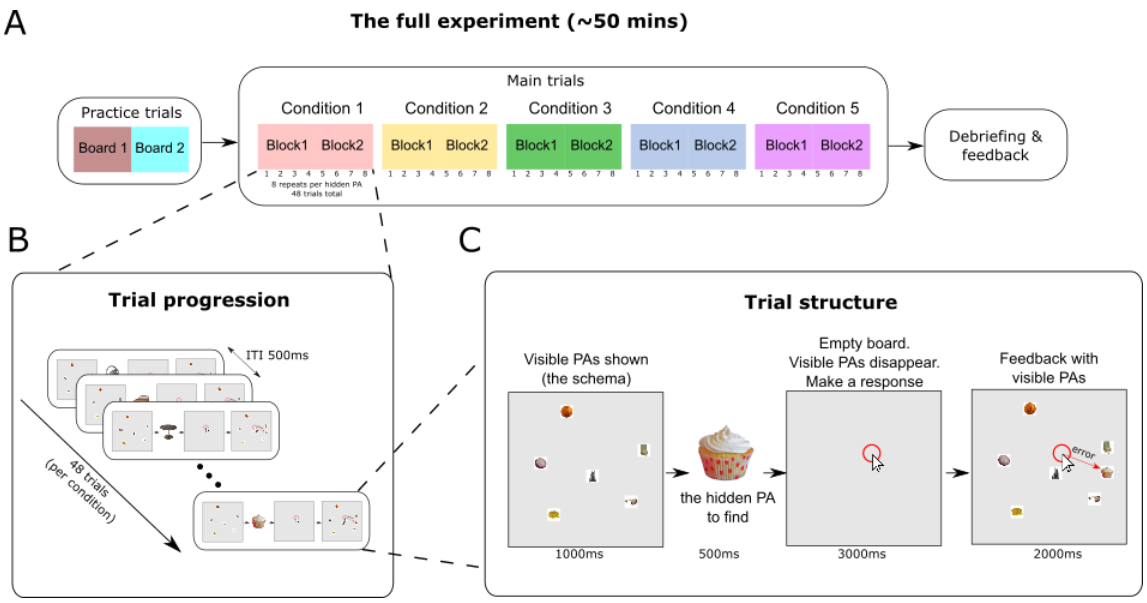


Figure 5.3: Experiment 1 design, trial progression and trial structure.
(A) Participants started with two practice rounds. During the main trials, participants progressed through 5 learning conditions, each broken up into two learning blocks. Each Hidden-PA was repeated 8 times within a condition. Finally, a feedback survey was given with debriefing. (B) Trial progression. Total of 48 trials occurred within each learning condition, with an ITI of 500ms. (C) Trial structure. Participants started by seeing the board with the Visible-PAs (or no PAs if in the No-Schema condition). Then, the board disappeared and one of the Hidden-PA images appeared as a target for that trial. Following this, an empty board appeared, and participants made a response using a mouse. Finally, feedback was given by displaying the correct location of the Hidden-PA and all the other Visible-PAs.

5.2.1.3.2 Learning conditions

The Hidden-PAs remained at consistent locations on the 5 boards throughout the experiment. The 5 learning conditions differed in the stability of the Visible-PAs:

1. Schema-Consistent (Schema-C) condition: the 6 Visible-PAs remained located at the same spot on the board across trials.
2. Schema-Landmark (Schema-L): Across trials, 4 of the 6 Visible-PAs randomly relocated anywhere on the board, while 2 of them always remained stable.
3. Schema-inconsistent (Schema-IC): Across trials, the Visible-PAs appeared at the same 6 locations on the board, but swapped places with each other on every trial.

4. Random: On each trial, all 6 of the Visible-PAs moved to new locations anywhere on the board.
5. No-schema: No Visible-PAs were on the board. The participants learned the locations of the Hidden-PAs only.

Prior to the beginning of each learning condition, the participants were given instructions which showed the empty board, were separately shown all 12 visible and Hidden-PAs, and were asked to name all of the PAs in a text box. For the No-schema condition, the participants named only the 6 Hidden-PA images. The participants were not told which condition they were in, i.e. whether the Visible-PAs would remain stable or not across trials.

5.2.1.3.3 PA arrangements

There were 5 pairs of arrangements of the visible and Hidden-PAs: A, B, C, D, and E. All arrangements had hidden and Visible-PA locations at least 1 row and column away from the borders of the board, to avoid ceiling effects. The assignment of learning conditions (1 through 5, as described above) and PA-arrangements (A through E) is shown by an example sequence of participants P1-P5:

P1 A1 B2 C3 D4 E5

P2 A2 B3 C4 D5 E1

P3 A3 B4 C5 D1 E2

P4 A4 B5 C1 D2 E3

P5 A5 B1 C2 D3 E4

Thus, while the order of conditions was rotated across participants, the order of PA-arrangements was fixed, such that the assignment of PA-arrangement to condition was rotated.

5.2.1.3.4 Near versus Far-PAs

For the Schema-C and Schema-L conditions, 2 of the Hidden-PAs (called *Near-PAs*) were located adjacent to Visible-PAs that remained stable across trials (Figure 5.2-A). These 2 hidden Near-PAs had 2 corresponding hidden *Far-PAs*, which were equally distant from the border of the board but which had no adjacent Visible-PAs.

5.2.1.3.5 Practice trials

Before the 5 learning conditions, participants did 2 rounds of practice trials with 2 different boards. On each board, the participants had to find 3 Hidden-PAs, each repeated twice. The first board contained 6 Visible-PAs which remained stable across all trials. Afterwards, the participants were told that some of the boards in the real experiment will have Visible-PAs that move around (such as in the Schema-L or Random condition), while others stay stable, and that they will now do practice trials with such a board. The second board involved 2 stable Visible-PAs with 4 moving randomly across the 6 trials. At the end of each practice round, the participants had an option to re-read instructions and re-do the practice trials.

5.2.1.3.6 Feedback and debriefing

At the end of the experiment, the participants were asked whether they noticed the differences between the five learning conditions, whether the Visible-PAs helped or hindered them in learning the Hidden-PA locations, and whether they had any additional feedback.

5.2.1.4 Quality checks

The participants were recruited from prolific.co with the following pre-screening criteria:

- Current country of residence: UK or Ireland.
- Age: 18-40
- Fluent languages: English
- Vision: normal or corrected-to-normal
- Approval rate on prolific: minimum 95%
- Minimum number of previous submissions: 2

A participant was excluded from data analysis if they failed any of the following post-experiment QC screenings:

- Every page of the instructions was looked at for at least 1 second.
- For each condition, total number of missed trials OR trials with RT < 350ms were not more than 20%.
- Instructions were understood and followed, as indicated in the feedback forms.
- No technical errors interfered with the study, as indicated in the feedback forms.

Additionally, we used the following performance-based exclusion criteria:

- A participant was excluded if the overall accuracy in the 2nd block was below the 95th percentile of their participant-specific permutation-based null distribution of accuracy scores. Such distribution was computed by randomizing the mapping between the "correct response label" on each trial (i.e. which Hidden-PA was prompted to be found) and the participant responses. The 120 "correct response labels" were shuffled while keeping the participant responses unchanged, which maintained any biases or trial-to-trial response-dependencies in participants' data. Mean accuracy was computed for each such permutation. A total of 10,000 permutations was performed for each participant.
- We used a standard non-parametric exclusion criterion based on the first and third quartiles (Q1 and Q3) and the interquartile range (IQR) A participant was excluded if the mean accuracy in the 2nd block across all the conditions was above the Q3 + 1.5xIQR or below Q1 - 1.5xIQR of group data.

5.2.1.5 Data analysis

The preregistration document specifying our experimental manipulations, planned primary and secondary analyses and power calculation can be found here: https://osf.io/2znw5. Any deviations from the preregistered plans are explained below.

We used Matlab R2020a (www.mathworks.com) and R RStudio (http://www.rstudio.com/) with R statistical software (R Core Team, 2022) for data preprocessing and analysis.

5.2.1.5.1 Dependent variables:

On each trial, we captured the Euclidean error between the mouse click and the centre of the Hidden-PA location. The main dependent variable for each learning condition was the average Block 2 accuracy, which was computed separately for all Hidden-PAs, the Near-PAs, and the Far-PAs. Data were log transformed for normalization purposes.

As a secondary dependent variable, we estimated the learning rate per condition per participant, by fitting an exponential function to the average error on all 8 repetitions of the Hidden-PAs (averaged across the 6 Hidden-PAs). We fit a 2-parameter and a 3-parameter model as described below, and used the AIC criterion to determine the winning model.

- The 3-parameter model formula: $y = b * (e^{-c*(t-1)} - 1) + a$, where y is the Euclidean error as above, *t* is the trial number, *c* is the learning rate, *(a – b)* is the asymptote (e.g, motor error even when participants know exactly where an associate is), and *a* is the intercept (when *t=1)*.
- The 2-parameter model formula: $y = a * e^{-c*(t-1)}$, where y is the error, *t* is the trial number, *c* is the learning rate, and *a* is the intercept.
- *a* and *(a – b)* were bounded between 0 and maximum possible error on the board. No bounds were be applied to the learning rate *c*.
- In cases where the learning rate *c* for a participant was below Q1 – 1.5xIQR or above Q3 + 1.5xIQR for group data for that learning condition, it was classified as an outlier and was replaced by the average learning rate for that learning condition.
- If a participant missed trials such that no data point could be calculated for the 1st repetition of the PAs, the participant's data were excluded.

5.2.1.5.2 Predicted outcomes and planned contrasts

Predictions for each of the three hypothesis we tested are presented in Figure 5.2 B and D. Our main hypothesis concerned the "local versus global" influence of elements, with the "location knowledge" and the "distraction hypotheses" as secondary comparisons.

5.2.1.5.2.1 The local versus global influence hypothesis

We hypothesized that if individual schema elements (i.e. the Visible-PAs) only have a local influence, that is no global network of connected knowledge exists:

- In the Schema-C condition, learning of the Near-PAs will be faster and better than Far-PAs, i.e. error will be smaller in Block 2.
- Performance on Schema-C Near-PAs will be similar to that for Schema-L Near-PAs.

If schemas exist as interconnected knowledge structures and globally influence learning:

- In the Schema-C condition, Near and Far-PAs will have similar performance.
- Schema-C Near-PAs will be learned better than Schema-L Near-PAs.

Therefore, we ran the following two contrasts:

1. Contrast 1: within Schema-C, Near-PA versus Far-PAs.

2. Contrast 2: Schema-C Near-PAs versus Schema-L Near-PAs.

The difference scores for Contrast 1 were subjected to a Bayesian one-sided, one-sample t-test, whereas we used a two-sided test for Contrast 2 in order to more appropriately support a possible absence of an effect[5].

5.2.1.5.2.2 The location knowledge hypothesis

The participants might extract a location knowledge in the Schema-IC condition, realising that the locations on the board remain stable even if images swap places. Such a schema might be enough to facilitate learning, in which case performance in Schema-IC would be comparable to Schema-C, and Schema-IC should display the Near versus Far-PA advantage. These two contrasts formed secondary comparisons for our experiment.

5.2.1.5.2.3 The distraction hypothesis

To test if randomly moving Visible-PAs on every trial might interfere with Hidden-PA learning, we compared overall learning in the Random condition to the No-Schema condition[6].

5.2.1.5.3 Sample size and power calculation

Similar to Chapter 2, we used a sequential Bayesian design with maximal N. The initial starting n was set to 20 participants. The stopping criteria for data acquisition were based on the two main contrasts for testing local versus global effects:

1. Contrast 1: Near- versus Far-PAs within Schema-C.
2. Contrast 2: Schema-C Near-PAs versus Schema-L Near-PAs.

If for both contrasts, the BFs exceeded the threshold of 6 (whether in support of H0 or H1), we stopped data collection. Batch size was set to 15, while maximum N was set to 110 valid participants. The maximum of N=110 was decided based on simulating "power" for supporting one of the two following possibilities:

[5] In our pre-registration, we planned to use a one-sided test for Contrast 2 as well. However, support for the null in a one-sided test could be due to absence of an effect or presence of an effect in the opposite direction. Therefore, to support a possible absence of an effect more appropriately, we report the results of a two-sided one-sample t-test for Contrast 2.

[6] This secondary comparison was not part of the original preregistered plan.

1. Scenario 1: existence of true medium sized effect (Cohen's $d_1 = 0.5$) for Contrast 1 and no effect ($d_2 = 0$) for Contrast 2. This would support the "local influence" hypothesis.
2. Scenario 2: no effect for Contrast 1 ($d_1 = 0$), and a medium sized effect for Contrast 2 ($d_2 = 0.5$). This would support the "global influence" hypothesis.

We performed 10,000 simulations of our Bayesian sequential design for the joint outcomes of Contrasts 1 and 2. The table below illustrates the frequencies of various outcomes from the simulations for Scenario 1, i.e. when $d_1 = 0.5$ and $d_2 = 0$:

For Contrast 1 supports:	**For Contrast 2 supports:**	**Percent of simulations:**
H1	H0	78.7%
H1	Undecided	18.9%
H1	H1	1.69%
Undecided	H0	0.36%
H0	H0	0.19%
Undecided	Undecided	0.19%
H0	H1	0.01%
H0	Undecided	<0.01%
Undecided	H1	<0.01%
Total:		100%

Thus, our procedure had 78.7% "power" to correctly support H1 for Contrast 1 when $d_1 = 0.5$ and to correctly support H0 for Contrast 2 when $d_2 = 0$. Note that since the two scenarios above are symmetrical, the procedure analogously had the same power to support H0 for Contrast 1 when $d_1 = 0$ and H1 for Contrast 2 when $d_2 = 0.5$.

5.2.2 Results

After acquiring a total of 65 valid participants, we reached the BF thresholds for both of our primary contrasts testing the local versus global effects. For Contrast 1, Near-PAs were better learned than Far-PAs within the Schema-C condition ($BF10 > 4.82 \times 10^3$), supporting a local effect. However, for Contrast 2, Near-PAs were better learned in the Schema-C condition than Schema-IC condition, supporting an additional global effect

(BF10 > 16.5). For a more complete description, including tests of our secondary hypotheses, we unpack the results below.

Analyses of learning rates corroborated the results with the Block 2 error variable, and are reported as supplementary material in the Appendices.

5.2.2.1 Analysis across all Hidden-PAs

Figure 5.4-A below shows learning (i.e. decrease in error) over all 6 Hidden-PAs in all 5 learning conditions, indicating that the participants successfully learned the task in all conditions.

Figure 5.4-B depicts mean Block 2 error over all Hidden-PAs for all conditions. Schema-C was superior to Schema-L (BF10 > 124), arguing that the distant stable Visible-PAs in the Schema-C condition could have still had a facilitatory effect on learning, supporting the global influence hypothesis. However, we also found a significant distraction effect, with the No-Schema condition outperforming the Random condition (BF10 > 77.3), meaning that randomly moving Visible-PAs had a negative effect on learning. This was corroborated with participant feedback, reporting such a distraction from moving PAs. Importantly, similar distraction effect could have played a role in the Schema-L condition, meaning that the Schema-C versus Schema-L difference could be due to such distraction instead of a facilitatory effect of the far-away stable Visible-PAs in the Schema-C condition. We try to disentangle this in the next experiment.

Schema-C also outperformed other conditions (Schema-C versus Random BF10 > 5.50×10^5; Schema-C versus No-Schema BF10 > 37), except for the Schema-IC condition (BF01 = 3.96). Thus, presence of stable landmarks aided in learning, whether these landmarks were picture-location associations (Schema-C) or just locations (Schema-IC).

5.2.2.2 Near versus Far-PA analysis

Figure 5.4-C shows performance separately for the Near and Far-PAs in each condition. As expected, Schema-C and Schema-L conditions showed large advantages for Near-PAs (BF10 > 4.73×10^4 and BF10 > 2.40×10^3, respectively), but so did the Schema-IC condition (BF10 > 29.7), indicating again that mere location stability was sufficient to exert an effect. As expected, Random and No-Schema conditions showed no differences for Near versus Far-PAs (BF01 > 13, BF01 > 9.45), as no stable landmarks were present (and demonstrating no confounding difference between Near and Far-PAs in terms of their locations). These results are in accordance with the local influence hypothesis,

showing significant benefits of being nearby a stable landmark in Schema-C and Schema-L conditions.

However, contrary to the local influence hypothesis, we found that the Near-PAs in Schema-C were learned better than those in Schema-L ($BF10 > 4.8 \times 10^3$), consistent with the global influence hypothesis that distant stable landmarks might still exert positive influence on the two Near-Hidden-PAs in Schema-C. As with the analysis of all Hidden-PAs above, an alternative reason for this difference could be the distracting effect of randomly moving Visible-PAs in the Schema-L condition.

Instead of contrasting Schema-C versus Schema-L, which is confounded by distraction, one could compare performance between Schema-C and No-Schema for Near and Far-PAs separately, to capture any local or global effects in absence of distraction[7]. If the Near-PAs of Schema-C outperform Near-PAs of No-Schema (which had no adjacent landmarks), this could be due to the local effects of adjacent Visible-PAs, but also additional distant effects of the far-away Visible-PAs in the Schema-C condition. However, if at the same time, we find no difference between Far-PAs in the Schema-C condition versus the Far-PAs of the No-Schema condition, this would exclude any beneficial distal effects of far-away Visible-PAs in the Schema-C condition. We found no difference between the Far-PAs of Schema-C versus No-Schema ($BF01 > 6.76$), but extreme support for Schema-C Near-PAs outperforming No-Schema Near-PAs ($BF10 > 10.5 \times 10^3$). This exploratory analysis supports the local influence hypothesis, while not being confounded by distraction effects, arguing that learning is enhanced by the presence of local landmarks, but that more distal landmarks have no beneficial effects.

[7] This analysis was not planned and is not part of the pre-registration. However, it is particularly suited to testing effects of far-away Visible-PAs without any distraction confounds, so we report it as exploratory analysis.

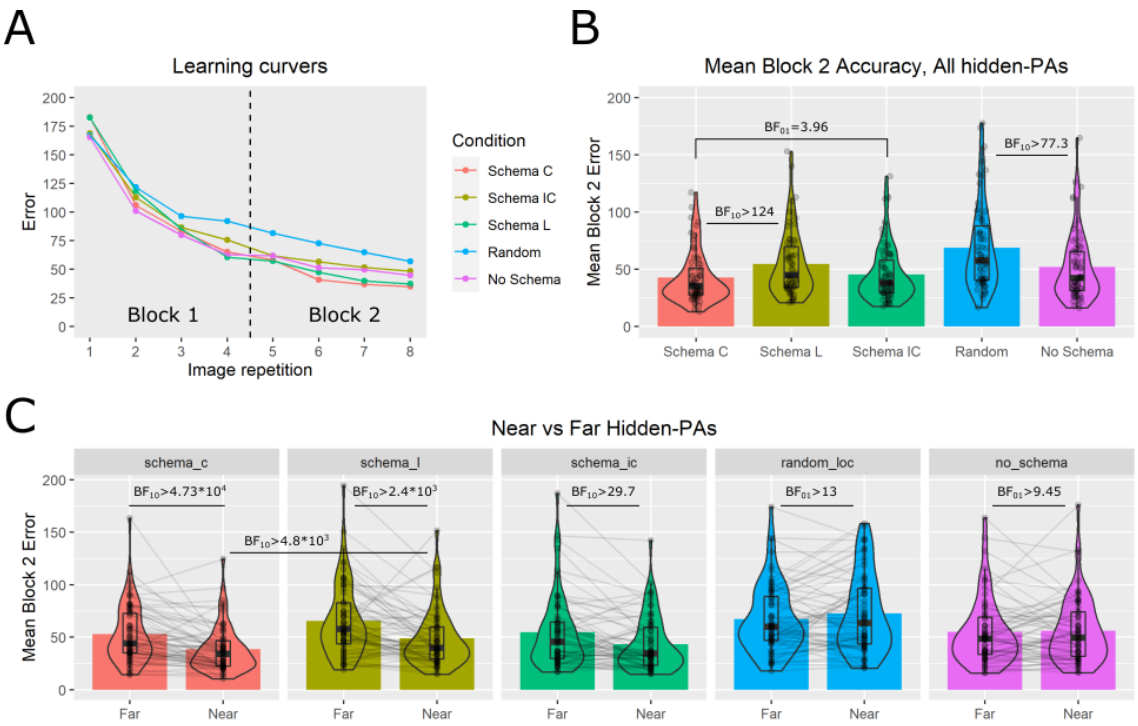


Figure 5.4: Experiment 1 results (compare Figure 5.2 for predictions).
(A) Averaging learning over all Hidden-PAs over all participants showed a steady decrease in error in all conditions. (B) Average Block 2 error across all Hidden-PAs for each condition. Height of the bars represent within-group mean error. Schema-C outperformed Schema-L, as predicted by the "local influence" hypothesis. Schema-C showed no difference from Schema-IC, as predicted by the "location knowledge" hypothesis. No-Schema outperformed Random, as predicted by the "distraction" hypothesis. (C) Near- versus Far-PA comparison within each condition. Schema-C, Schema-L and Schema-IC all showed advantage for Near-PAs, indicating that being located adjacent to a landmark had a beneficial effect on learning. This pattern is consistent with predictions of the "local influence" hypothesis and "location knowledge" hypothesis. However, Near-PAs of Schema-C outperformed Near-PAs of Schema-L, which is predicted by both the "global influence" hypothesis and the "distraction" hypothesis.

5.2.3 Discussion

In accordance with the local influence hypothesis, Experiment 1 confirmed our a priori expectation that in the Schema-C condition, those Hidden-PAs that had a stable neighbouring Visible-PA were learned better than those without a stable neighbouring Visible-PA. However, the second prediction of the local influence hypothesis was

refuted: we did not find that Schema-C Near-PAs were learned comparably to those in the Schema-L condition, but rather that they were learned better in the Schema-C condition. One explanation for this is an additional global effect in the Schema-C condition, i.e. the presence of four other stable PAs, even if far from the Near-PAs being learned, additionally helped memory in the Schema-C condition (e.g., through an abstract schema that encoded locations of Visible-PAs relative to each other). However, there is an alternative explanation for this second finding: the movement of the other four Visible-PAs in the Schema-L condition was distracting to participants, impeding their learning. This was corroborated by feedback from the participants, and further supported by the additional finding of worse performance in the Random condition, which had 6 randomly changing Visible-PAs, than in the No-schema condition, which did not have any Visible-PAs. Thus, according to this alternative "distraction" hypothesis, Schema-C and Schema-L differ not only in possible stronger global schema structure (6 stable Visible-PAs for Schema-C versus only 2 in Schema-L), but also in the number of randomly moving items (0 random Visible-PAs in Schema-C versus 4 in Schema-L).

Thus, our pre-registered analyses could not confirm presence or absence of global effects. However, in an exploratory comparison, we tested for global effects without distraction confounds by separately comparing the Near and Far-PAs of the Schema-C condition to those of the No-Schema condition. We found no difference between these conditions for the Far-PAs, arguing against any distal effects of stable Visible-PAs, while at the same time we found strong evidence for the Near-PAs of the Schema-C condition outperforming those in the No-Schema condition, arguing in support of local positive influences of landmarks. Given the exploratory nature of this result, we pre-registered Experiment 2 to confirm the presence or absence of the two potential influences on learning: global influences versus distraction.

Through our secondary analysis, we also compared the Schema-C with Schema-IC, expecting the former to outperform the latter. Although the Bayes Factor in support for the null did not exceed the threshold of 6, the data suggested no difference between these two conditions. Furthermore, Schema-IC condition also displayed the local facilitatory effect on learning of Hidden-PAs. It appears that, simply having stable locations for Visible-PA images is enough to allow them to become landmarks, even if images at those locations swap places on every trial. This contrasts with previous schema experiments (Guo & Yang, 2020; Tse et al., 2007; van Buuren et al., 2014), where Schema-C was consistently found to be superior to Schema-IC. This discrepancy likely reflects one or

more important differences in design between our paradigm and previous paradigms. For example, we had no separate schema learning stage – our participants saw the Visible-PA structure at the beginning of each trial – which may be sufficient to learn the locations, but not be sufficient to learn the associated images, which may require more prolonged training.

5.3 Experiment 2

In the second experiment, we attempted to confirm the absence of global facilitatory effects of stable schema elements (i.e. fixed Visible-PAs) and the distracting effects of randomly moving Visible-PAs. Additionally, we tried to replicate the Near versus Far finding of the Schema-C condition from Experiment 1. As in Experiment 1, participants had to find 6 Hidden-PAs on different boards. We designed four learning conditions, depicted on Figure 5.5-A.

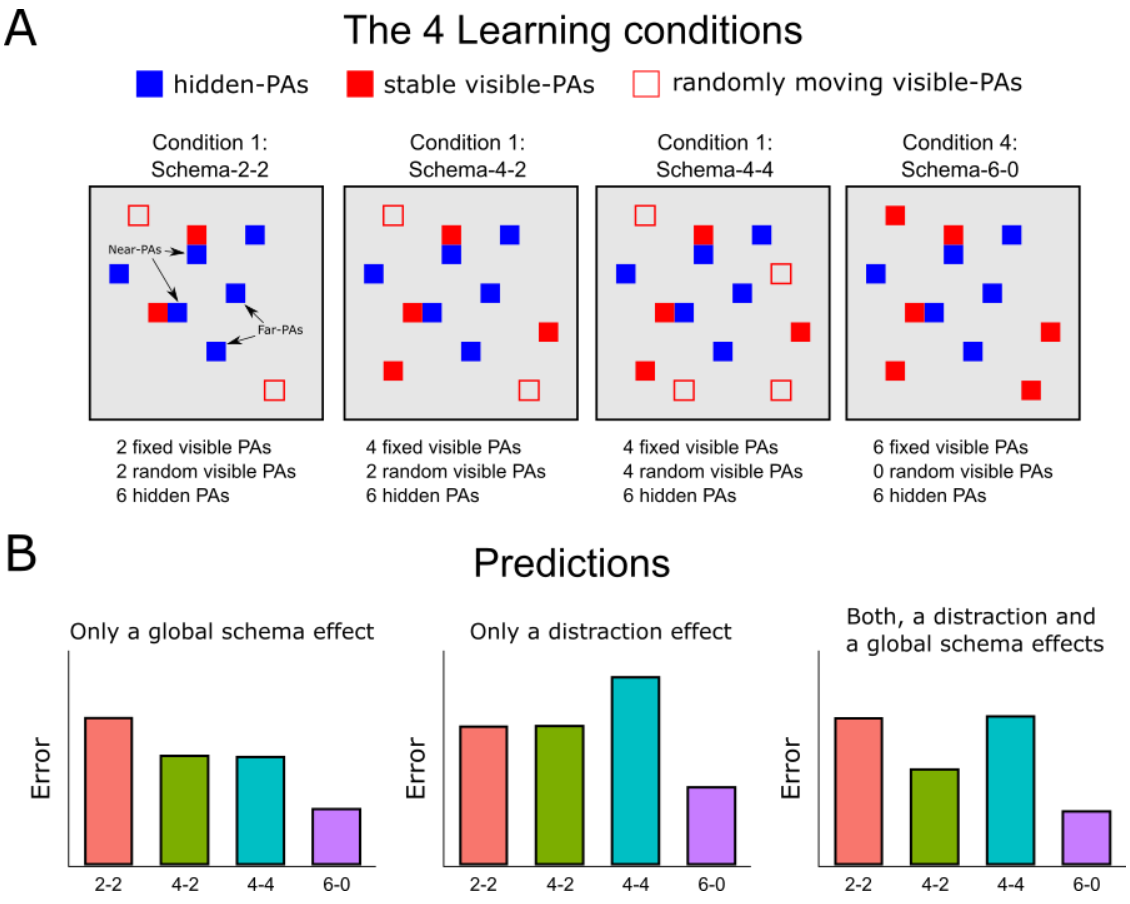


Figure 5.5: Experiment 2 learning conditions and predictions.

(A) The 4 learning conditions, involving varying numbers of fixed and randomly moving Visible-PAs. (B) Predictions for the "global influence" hypothesis, the "distraction" hypotheses, and a combination of the two.

5.3.1 Isolating the global facilitatory effect of fixed Visible-PAs

According to the global influences hypothesis, having additional fixed Visible-PAs helps in learning Hidden-PA locations, even if these Visible-PAs are not adjacent to Hidden-PAs. Thus, two conditions that differ *only* in the number of far-away fixed Visible-PAs should show a difference in learning. To this end, we compared the following two conditions. In the Schema-2-2 condition, 2 Visible-PAs remained stable while 2 others randomly moved on every trial. The 2 stable Visible-PAs were adjacent to Hidden-PAs (i.e. the Near-PAs). In the Schema-4-2 condition, 4 Visible-PAs remained stable while 2 randomly moved. Thus, the only difference between Schema-4-2 and Schema-2-2 conditions was in the extra 2 Visible-PAs in the Schema-4-2 condition. We expected better performance in the Schema-4-2 condition, and this comparison formed the first of the two primary contrasts for Experiment 2.

5.3.2 Isolating the distracting effect of random Visible-PAs

Our third learning condition, the Schema-4-4, differed from Schema-4-2 only by having 2 extra randomly moving Visible-PAs. Thus, if random movement of PAs on every trial has a negative effect on learning, we would expect worse performance in Schema-4-4 versus Schema-4-2. This comparison formed the second primary contrast for Experiment 2.

5.3.3 Replicating the Near-Far difference

We included a 4th condition that was identical to the Schema-C condition from Experiment 1, called Schema-6-0. All 6 Visible-PAs remained stable across trials. 2 Hidden-PAs were adjacent to visible PAs while 2 were far away. This allowed us to replicate the Near-vs-Far finding from Experiment 1.

5.3.4 Methods

5.3.4.1 Participants

139 healthy young adult participants were recruited (66 females) from the prolific.co platform, aged 18-40 (M = 30.6, SD = 6), and paid £6/hour for their time, according to the Cambridge Psychology Research Ethics Committee protocol PRE.2020.018. Of these, 116 (55 females, 30 males) aged 18-40 (M = 30.66, SD = 5.89) passed the final quality and performance checks (see the *Quality checks* section below) to be included in the data analysis.

5.3.4.2 Stimuli

The 2D board size parameters and PA images chosen were same as for Experiment 1.

5.3.4.3 Task design and procedure

5.3.4.3.1 Trial structure:

The trial structure was identical to that of Experiment 1.

5.3.4.3.2 Learning conditions

There were 4 learning conditions (see Figure 5.5-A). As in Experiment 1, the participants had to find 6 Hidden-PAs on each of the boards. The conditions differed by the number and type of Visible-PAs shown at the beginning of each trial:

1. Schema-2-2: 2 Visible-PAs remained fixed across trials while 2 others moved anywhere on the board.
2. Schema-4-2: 4 Visible-PAs remained fixed across trials while 2 others moved anywhere on the board.
3. Schema-4-4: 4 Visible-PAs remained fixed across trials while 4 others moved anywhere on the board.
4. Schema-6-0: All the 6 Visible-PAs remained fixed across trials. This condition was analogous to the first condition in Experiment 1 and was included to replicate the result with Near versus Far-PA difference.

As with Experiment 1, learning within each condition was broken up into two blocks, with each Hidden-PA repeated 8 times. A small break was given between the blocks and learning conditions. Prior to the beginning of each learning condition, the participants were given instructions which showed the empty board where they would have to learn the Hidden-PA locations, were separately shown all visible and Hidden-PAs, and were asked to name all of the PAs in a text box.

5.3.4.3.3 PA arrangements

There were 4 pairs of arrangements of the visible and Hidden-PAs: A, B, C, and D. The assignment of learning conditions (1 through 4, as described above) and PA-arrangements (A through D) is shown by an example sequence of participants P1-P4 (analogous to Experiment 1):

P1 A1 B2 C3 D4

P2 A2 B3 C4 D1

P3 A3 B4 C1 D2

P4 A4 B1 C2 D3

5.3.4.3.4 Near versus Far-PAs

For each of the four conditions, we included two Near-PAs and two Far-PAs, similar to Experiment 1. This allowed for a replication of the Near-Far difference result of Experiment 1.

5.3.4.3.5 Practice trials

Similar to Experiment 1, practice trials involved finding Hidden-PAs on two boards, one with stable Visible-PAs and one with some of the Visible-PAs moving randomly. At the end of each practice round, the participants had an option to re-read instructions and re-do the practice trials.

5.3.4.3.6 Debriefing

At the end of the experiment, the participants were asked whether they noticed the differences between the learning conditions, whether the Visible-PAs helped or hindered them in learning the Hidden-PA locations, and whether they had any additional feedback.

5.3.4.4 Quality checks

A participant was excluded if they failed any of the following post-experiment QC screenings:

- For each condition, check that the total number of missed trials OR trials with RT < 350ms are not more than 20%.
- Each break did not last for more than 10 minutes.
- Indication during the debriefing survey of not having understood the instructions or failed to have followed them.
- Indication during the debriefing survey of having encountered any technical error that interfered with the study.
- Indication that they had display issues that interfered with proper conduction of the study, such as having to scroll to see the full board before making a response.

Performance-based exclusion criteria were the same as for Experiment 1.

5.3.4.5 Data analysis

The preregistration document specifying our experimental manipulations, planned primary and secondary analyses and power calculation can be found here: https://osf.io/9xc3w. Any deviations from the preregistered plans are explained below.

5.3.4.5.1 Dependent variables:

As with Experiment 1, the main dependent variable was the average Block 2 error. Data were log transformed to satisfy normality.

As with Experiment 1, as a secondary dependent variable we estimated the learning rates per condition and per participant. Only a 2-parameter model was used:

- Model formula: $y = a * e^{-c*(t-1)}$ where y is the error, t is the trial number, c is the learning rate, and a is the intercept.
- a and b will were bounded between 0 and maximum possible error, while c had an upper bound of 4.09 determined through the procedure described below.
- Determining the upper bound of learning rate c: We simulated a hypothetical learning data across the 8 repetitions, starting with the maximum possible error (~620 pixels) on the 1st repetition and immediately dropping to an error of 10 pixels (to account for the motor error in responding with a mouse) on the 2nd repetition. Fitting the 2-parameter model to such learning data resulted in the learning parameter estimate of c=4.09. This indicates that larger learning curve estimates will not provide substantially better fits to even a perfect learner scenario, but they would introduce skewness in the distribution of learning estimates.
- If a participant missed trials such that no data point could be calculated for the 1st repetition of the PAs, the participant's data was excluded.

5.3.4.5.2 Predicted outcomes and planned contrasts:

The analysis and predictions below pertain to the combined Block 2 error rates on the 4 Hidden-PAs in the centre of the board: the 2 Near-PAs and the 2 Far-PAs. The other 2 Hidden-PAs that were close to the border were not analysed, since data from Experiment 1 showed ceiling effects on these PAs.

Figure 5.5-B depicts the relevant possible outcomes under different hypotheses. We predicted that the number of fixed items will have a facilitatory effect while the number

of random items will have a distracting effect (i.e. the 3rd scenario in Figure 5.5). Thus, we predicted the following key inequalities will hold for the Block 2 error in the following conditions:

- Contrast 1: Schema-2-2 > Schema-4-2, since Schema-4-2 has more fixed items.
- Contrast 2: Schema-4-4 > Schema-4-2, since Schema-4-4 has more random items.

The difference scores were subjected to a Bayesian two-sided one-sample t-test.

5.3.4.5.3 Sample size and power calculation

As in Experiment 1, a sequential Bayesian design with maximal N was used. The initial n was set to 20, and BF_{10} and BF_{01} were calculated for the two contrasts outlined above. If for both contrasts, the Bayes factors exceeded the threshold of 6 (whether in support of H0 or H1) we stopped data collection. Batch size was set to 16, while maximum N was set to 116 valid participants.

The maximum of N=116 was decided based on simulating "power" for supporting one of the three following possibilities (as schematically depicted on Figure 5.5):

1. Scenario 1: existence of a true medium sized effect (Cohen's $d_1 = 0.5$) for Contrast 1 and no effect ($d_2 = 0$) for Contrast 2. This would support the hypothesis that only the fixed landmarks have a positive influence.
2. Scenario 2: no effect for Contrast 1 ($d_1 = 0$), and a medium sized effect for Contrast 2 ($d_2 = 0.5$). This would support the hypothesis that only the random landmarks have a distraction effect.
3. Scenario 3: existence of true medium sized effect for both Contrast 1 and Contrast 2 (i.e. $d_1 = 0.5$ & $d_2 = 0.5$).

For each of the above scenarios, we performed 10,000 simulations of our Bayesian sequential design. The table below illustrates the frequencies of various outcomes from the simulations for Scenario 1, i.e. when $d_1 = 0.5$ and $d_2 = 0$. Note that Scenario 1 and 2 above are symmetrical, so our procedure has the same power to make the correct inference of existence of effects in both scenarios.

If Scenario 1, i.e. when $d_1 = 0.5$ and $d_2 = 0$:

For contrast 1 supports:	**For contrast 2 supports:**	**Percent of simulations:**
H1	H0	80.0%

H1	Undecided	17.6%
H1	H1	1.66%
Undecided	H1	0.01%
Undecided	H0	0.25%
H0	H0	<0.01%
Undecided	Undecided	0.39%
H0	H1	<0.01%
H0	Undecided	<0.01%
Total:		100%

Thus, we had 80% "power" to correctly support H1 for Contrast 1 when $d_1 = 0.5$ and to correctly support H0 for Contrast 2 when $d_2 = 0$. Likewise, we had 80% "power" to correctly support H0 for Contrast 1 when $d_1 = 0$ and to correctly support H1 for Contrast 2 when $d_2 = 0.5$.

If Scenario 3, i.e. $d_1 = 0.5$ & $d_2 = 0.5$:

For contrast 1 supports:	**For contrast 2 supports:**	**Percent of simulations:**
H1	H1	98.7%
H1	Undecided	0.6%
Undecided	H1	0.45%
H1	H0	0.14%
H0	H1	0.1%
Undecided	Undecided	0.01%
H0	Undecided	<0.01%
H0	H0	<0.01%
Undecided	H0	<0.01%
Total:		100%

Thus, we had 99% "power" to correctly support H1 for Contrast 1 when $d_1 = 0.5$ and H1 for Contrast 2 when $d_2 = 0.5$.

5.3.5 Results

5.3.5.1 No effect of either extra stable Visible-PAs or extra randomly moving Visible-PAs

As planned in our pre-registration document, the analysis reported below was run only on the 4 Hidden-PAs on each board (i.e. the Near and Far-PAs combined).

Despite reaching our maximum N=116 valid participants, we obtained evidence in support of only one of our planned comparisons. Analysis of Block 2 scores (Figure 5.6-B) showed no difference between Schema-2-2 and Schema-4-2 conditions (two-sided Bayesian paired t-test, BF01 > 9.6), in accordance with the local influence hypothesis. We did not find supporting evidence for the distraction effect when comparing Schema-4-4 with Schema-4-2, since the Bayes Factor was inconclusive (BF01 = 1.01) despite Schema-4-4 having largest error numerically (Figure 5.6-B). Analysis of learning rates for Schema-4-4 versus Schema-4-2 showed anecdotal evidence for an absence of an effect (BF01 > 5.6), despite being numerically lowest in the Schema-4-4 condition (supplementary Supplementary Figure 8.8). As an additional exploratory analysis[8], we compared the conditions for the overall error across both blocks, finding Schema-4-4 to have the largest error numerically, but the BF10 in support of existence of an effect was anecdotal at 3.34 (supplementary Supplementary Figure 8.9).

Thus, having additional Visible-PAs in Schema-4-2 did not help as compared to Schema-2-2. We did not obtain conclusive evidence that having two additional randomly moving Visible-PAs in Schema-4-4 as compared to Schema-4-2 hurts performance in Block 2, although exploratory analysis of combined Blocks 1 and 2 Error rates showed anecdotal evidence in support of a difference. This is in contrast with the learning rate analysis, which showed anecdotal evidence in favour of an absence of an effect.

5.3.5.2 Near-PAs show benefit over Far-PAs as before

The Schema-C condition in Experiment 1 showed better performance for Near-PAs than Far-PAs. Here, in an equivalent condition of Schema-6-0, we confirmed this effect (one-

[8] Not pre-registered.

sided BF10 > $1.26x10^9$). Figure 5.6-C shows that this effect was present for all four of our learning conditions, confirming the robustness of the finding.

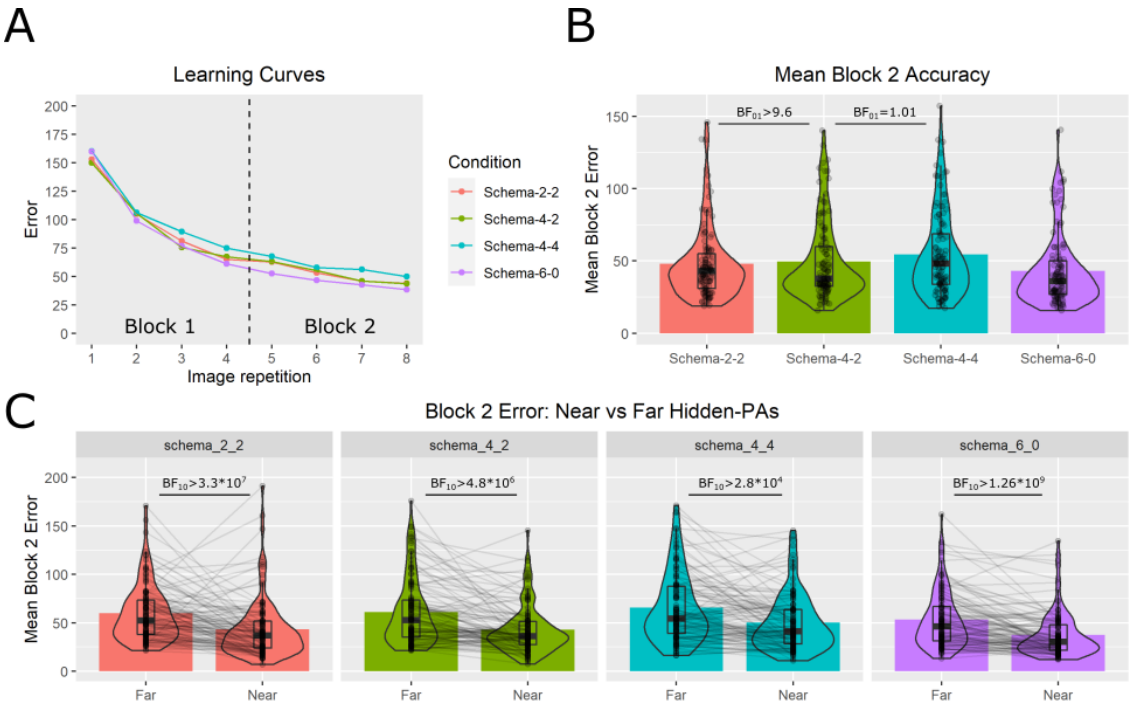


Figure 5.6: Experiment 2 results. Compare to Figure 5.5 for predictions.
(A) Average error across the 4 middle Hidden-PAs showed consistent learning for all conditions. (B) Average Block 2 accuracy for the 4 middle Hidden-PAs, for each condition. Heights of the bars represent within-group mean error. (C) Average Block 2 error between the Near-PAs and the Far-PAs, separately for each condition.

5.3.6 Discussion

In Experiment 2, we showed that having additional stable Visible-PAs do not influence performance if these Visible-PAs are not directly neighbouring the to-be-learned Hidden-PAs. We also replicated the Near-Far difference of Schema-C condition in Experiment 1; indeed, the Near advantage was found in all learning conditions. Regarding the distraction hypothesis, we found conflicting evidence when analysing averaged error versus learning rates. Although learning rate analysis showed anecdotal evidence for a lack of an effect and the data from our main dependent variable of Block 2 Error was inconclusive, the combined error across both blocks anecdotally supported presence of an effect, suggesting that distraction might have impacted early stages of learning too (i.e. in Block 1; see Figure 5.6-A). Existence of an effect would indicate that performance was influenced by distraction from moving Visible-PAs, which would be in line with the results of Experiment 1 where performance for Near-PAs in the Schema-L condition was

worse than for Near-PAs in the Schema-C condition, and performance in the Random condition was worse than the No-Schema condition. We further elaborate on our results from the two experiments in the General Discussion.

5.4 General Discussion

In this chapter, we present a brief overview of the recent literature showing how schema might influence learning of new spatial locations, and raise an important open question regarding the nature of this influence. Several previous rodent and human studies have examined schema effects by experimentally inducing schema by contrasting learning in environments in which a number of stimulus-location associations are consistent across trials, relative to when they are inconsistent. Schemas have consisted of paired-associate (PA) knowledge, such as rats learning flavour-place associations in a 2D arena to find food (e.g. Tse et al., 2007), or humans learning image-location associations on a 2D grid board on a computer screen (Guo & Yang, 2020, 2022; van Buuren et al., 2014). These studies have found that, in a condition where stimulus-location mappings remain consistent across training, learning of novel PAs is accelerated, compared to a condition where the stimulus-location mapping gets shuffled. Such shuffling involves swapping of associations, such that the same locations are used, but are paired with a different stimulus, thus preventing establishment of a schema defined by a number of PAs.

The interpretation of these results has been that the consistent PAs form a schema as an interconnected network of knowledge, in terms of spatial relationships between PAs, which is used as a scaffold to incorporate new information (Ghosh & Gilboa, 2014; Gilboa & Marlatte, 2017; McClelland, 2013; van Kesteren et al., 2012, 2013). However, an unexamined alternative explanation is that faster learning of any new piece of information is facilitated simply by learning its relationship to the nearest PA (or "landmark"), regardless of any of the other PAs. That is, performance might be benefited by a number of independent memories for each PA, without any representation of the relationships between PAs (i.e. without a schema) – what we have called a "local" versus "global" effect. Across 2 experiments, we tried to disentangle such local versus global influences by adapting previous human schema tasks, such that some new paired-associates (Near-PAs) were directly adjacent to schema elements, while others (Far-PAs) had no neighbouring landmarks. Additionally, we got rid of a separate schema-learning stage, instead displaying the schema items (image-location associations) on the board at

the beginning of each trial, making them directly available for the participants to scaffold their new knowledge.

In both Experiment 1 and 2, we found that to-be-learned items that were directly adjacent to schema elements (Near-PAs) were learned faster than those that had no such neighbouring landmarks (Far-PAs). Although this is consistent with the local influence hypothesis, the far-away schema elements could still have exerted an additional, beneficial global effect from the distance. We reasoned that such a distal effect should disappear if the far-away schema elements are not stable, but change locations on every trial. We implemented this in a different learning condition of Experiment 1 (Schema-L), where only two schema elements remained stable, while the rest randomly moved. We found that learning of Hidden-PAs directly adjacent to the two stable schema elements was worse in this condition than in the original Schema-C condition with stable far-away schema elements, supporting presence of some distal influence of these far-away stable items in the Schema-C condition.

An alternative explanation of the above result, however, could be that learning with only 2 stable landmarks and 4 moving ones could have suffered from a distracting effect of the moving items. This distraction effect was independently confirmed with a comparison of two control conditions, one with all the schema items moving on every trial (the Random condition) and one with no schema items on the board (No-Schema). Learning in the latter environment was better than in the former, indicating that moving images had a negative influence on performance.

What is more, exploratory analysis further supported absence of global effects. We separately compared the Far-PAs of the Schema-C condition to those of the No-Schema condition, which controlled for any distraction with the No-Schema condition that had no Visible-PAs. This exploratory contrast showed no difference in performance, indicating that the extra far-away Visible-PAs in the Schema-C condition did not exert any beneficial effect on learning. What is more, in a strong support for the local influence hypothesis, Near-PAs of Schema-C outperformed the Near-PAs of No-Schema, showing that adjacent landmarks facilitated learning.

To summarise Experiment 1, we found local beneficial effects of being next to a landmark element. Through our pre-registered analysis, we could not exclude the presence of global effects due to confounds of distraction in our crucial contrasts. However, exploratory comparisons showed that, while randomly moving PAs do distract compared to when no

PAs are present on the board (Random vs No-Schema comparison), having stable PAs does not exert any benefit from the distance (Schema-C vs No-Schema comparison).

In Experiment 2, we tried to confirm absence of global effects and presence of distraction effects by designing pairs of conditions that were identical except for either in number of far-away fixed landmarks (Schema-2-2 versus Schema-4-2 conditions) or in number of randomly moving landmarks (Schema-4-2 versus Schema-4-4 conditions). We found that extra distant landmarks did not induce better learning, arguing against the global influence hypothesis. This would imply that, in Experiment 1, the better learning of Near-PAs in Schema-C condition compared to Schema-L was not due to extra far-away schema elements in Schema-C, but was because of negative effects of extra randomly moving items in Schema-L. However, this distraction hypothesis was not confirmed in our second contrast of Experiment 2, where we failed to obtain decisive evidence that extra randomly moving items hinder learning. Although comparison of learning rates showed anecdotal support for the null, i.e. no difference caused by extra randomly moving items ($BF01 > 5.6$), the Bayes Factor for our main dependent variable of average Block 2 error remained moot ($BF10 = 1.01$).

A possible explanation of these discrepant results is that, while a distraction effect exists, as shown by No-Schema versus Random comparison in Experiment 1, the manipulation in Experiment 2 was not strong enough to elicit it. While in Experiment 1, the No-Schema versus Random conditions differed by 6 randomly moving items and the Schema-C versus Schema-L conditions differed in 4 randomly moving items, the Experiment 2 Schema-4-2 and Schema-4-4 conditions differed only by 2 extra randomly moving items. It is possible that the 2 extra random items had a small negative impact (much smaller than our assumed effect size of $d=0.5$ used to power the experiment), that was undetectable even after reaching our maximum N, explaining the indecisive Bayes Factor. This is supported by our exploratory finding that the overall error across both blocks showed anecdotal evidence in support of a distraction effect ($BF10 > 3.34$). Although this conflicts with learning rate analysis that showed anecdotal support for no distraction ($BF01 > 5.6$), it is possible that the learning rate (i.e. learning function) did not capture the distraction effect properly.

One important difference between our paradigm and those of previous studies concerns the absence of a separate schema-learning stage. This design choice was driven by pragmatic goals of having a fast and an efficient (single-session) study for online testing (since multi-session studies are less reliable in online testing). However, by directly

presenting the schema elements at the beginning of every trial, we might have distracted the participants in those conditions where these elements randomly moved. Thus, re-introducing a separate schema-learning stage could be one way for future experiments to control for the confounding effects of distraction, while retaining our Near versus Far manipulations to test for local versus global effects.

This difference in not having a schema-learning stage could explain the final noteworthy result in our study. In Experiment 1, learning in the consistent schema condition Schema-C was not different from learning in the Schema-IC condition where the picture locations remained stable, but the images swapped places on every trial. This means that, in our paradigm, simple knowledge of PA locations while ignoring the image identities was sufficient to drive learning, with no additional benefit gained from the stable image-location pairings in the Schema-C condition. This was supported by the finding that, similar to the Schema-C condition, this knowledge of PA locations in the Schema-IC condition was enough to exert a large Near-Far difference, such that learning of paired-associates with adjacent PA locations was faster than of those without such neighbouring landmarks. These results conflict with those from previous studies that have found large differences in learning between such consistent and inconsistent schema conditions (Guo & Yang, 2020, 2022; Tse et al., 2007; van Buuren et al., 2014), which is again potentially explained by a lack of a separate schema-learning stage in our design. It is possible that, on short time-scales, only the location information is encoded ignoring the image identity, while if training stretches over a long period, the stability of image-location pairings exerts an additional benefit over and beyond the mere location knowledge. Future studies with a separate learning stage should also compare Near vs Far performance across conditions with consistent vs inconsistent image-location mappings. This would test whether, over time, a benefit of consistent image-locations exerts a larger local influence on learning of new paired-associates, compared with only location knowledge.

5.5 Conclusion

Research on schemas has long demonstrated their beneficial effects on learning of consistent information. Exploration of the precise neurocomputational mechanisms underlying these psychological processes was initiated with rodent and human paradigms, where schemas were experimentally induced and subsequent learning was examined. We have argued that these paradigms have not sufficiently shown that the elements of purported schemas are encoded as an interconnected network of knowledge. Instead, it is

possible that each element is independently encoded and facilitates learning of consistent information only within its local neighbourhood. We presented two experiments that have attempted to disentangle such local versus global effects of associative knowledge elements, and found that learning is faster when close-by elements act as landmarks, while not finding much evidence to support that far-away elements still exert facilitatory effect from a distance.

6 DISCUSSION

6.1 Discussion

In this thesis, we have addressed the question of knowledge representation by (i) testing the proposed parallels between organisation of conceptual and spatial knowledge and (ii) developing experimental paradigms for examining spatial and non-spatial schemas. Adopting Marr's (1982) approach of analysing cognitive phenomena at various levels, our experiments have operated at the algorithmic level, elucidating the format in which concepts and other higher-knowledge structures are encoded. In this final discussion chapter, we summarise our findings in combination with the prior literature in order to highlight key conclusions and important open questions.

6.1.1 Geometric models of conceptual spaces

In the first two chapters, we tested the validity of geometric theories (Gärdenfors, 2000) which dominated the study of conceptual representation during the middle of the 20th century. These theories explained similarity judgments by proposing a simple and elegant distance formula (Equation 1.1), benefited from powerful visualisation tools such as MDS, and predicted behaviour on various cognitive tasks to an impressive level (e.g. Nosofsky, 1985a). Recently, formulation of concepts as regions in a multi-dimensional space has been supported by neural evidence, finding parallels between brain activity during navigation in physical spaces and "navigation" in conceptual spaces (e.g. Constantinescu et al., 2016). However, as we pointed out, an older line of behavioural work challenged the validity of geometric models, showing that similarity judgments often violated fundamental geometric axioms of minimality, symmetry, the triangle inequality and segmental additivity (Tversky, 1977; Tversky & Gati, 1982). This led Tversky (1977) to propose a rival algorithmic-level theory based on feature-sets, together with a formal contrast model for similarity calculations (Equation 1.2). Other theorists, however, proposed augmented geometric models to account for the axiomatic violations. In Chapter 2, we used similarity judgments in a one-dimensional stimulus space to test prediction of one such augmented model – the distance-density model (Krumhansl, 1978) – but failed to find supporting evidence. In Chapter 3, we used similarity judgments and

ideal observer simulations in two-dimensional stimulus spaces to find that violations of geometric axioms depended on the type of stimuli, and we discussed the ability of another type of augmented model – the attention-weighted geometric model (Gärdenfors, 2000) – to explain these data.

The distance-density model attempted to explain violations of the symmetry axiom documented by Tversky (1977) by incorporation of density in the geometric similarity calculation. A basic prediction of such a model is that increased density should stretch the psychological space, decreasing similarities between exemplars. Compared to previous studies that failed to find effects of density on similarity (Corter, 1987), we tested this basic prediction with a stronger density manipulation and a more powerful experimental design. Furthermore, we manipulated the neighbourhood density of certain exemplars instead of simply increasing their presentation frequency (as done by Polk et al. 2002). We have argued that changes in presentation frequency can be interpreted as changes in salience, and salience-induced asymmetries can be explained by Tversky's feature-based model as well. Despite these improvements, we were unable to detect any changes in similarities as a result of increased neighbourhood density. However, given that density manipulation was done on a short time-scale, and that our Bayes Factor did not reach our threshold for supporting the null, we recommend that future studies employ larger pools of participants and longer density manipulations to more definitively exclude effects of density on similarity.

These results argue that the distance-density model is an unlikely candidate for rescuing geometric theories in face of violations of the symmetry axiom. However, attention-weighted geometric models can explain asymmetries by suggesting that the order of item presentation during directional similarity judgments causes redistribution of attentional weights placed on activated dimensions, resulting in a different final distance computation. In Chapter 3, we tested various two-dimensional stimulus spaces for adherence to two other requirements of geometric models – segmental additivity and triangle inequality – and discussed applicability of the attention-weighted geometric model as well, as described below.

Our experiments were inspired by the seminal study of Tversky and Gati (1982), where they developed a test for the triangle inequality using only ordinal data from pair-wise similarity judgments, and went on to show that various 2D perceptual and conceptual spaces violated this ordinal triangle inequality. In our study, we found similar violations in two of our artificial stimulus spaces defined by psychologically separable quantitative

and qualitative dimensions, one of which was an adaptation of the 2D "square circle" space used by Theves et al. (2019). Furthermore, using ideal observer simulations, we showed that ordinal triangle inequality violations can be caused either by inter-dimensional superadditivity, i.e. when distances across dimensions combine into a total distance that is larger than their mathematical sum, or by intra-dimensional subtractivity, that is when smaller distances along a single dimension do not add up to a larger distance, thus violating segmental additivity. Importantly, we pointed out that, while inter-dimensional superadditivity could be explained by the attention-weighted geometric model, intra-dimensional subtractivity could not. Although some of our analyses indicated that our data were not inter-dimensionally superadditive, which would leave intra-dimensional subtractivity as the only explanation for ordinal triangle inequality violations, we could not definitively conclude this. However, in combination with the prior literature finding non-linear mapping between physical and psychological distances in various types of spaces (Fechner, 1860; Houston & Shearer, 1930; Weber, 1851), we argued that intra-dimensional subtractivity is the likely explanation for our data as well. We proposed that subsequent experiments might benefit from using eye-tracking or other methods to independently measure attention variation, in order to better test the validity of attention-weighted geometric model.

We also found that two other stimulus spaces – naturalistic bird stimuli defined by quantitative dimensions, one of which was adapted from Constantinescu et al. (2016) – did not show violations of ordinal triangle inequality. This was surprising as we expected such quantitative dimensions to be perceived as separable and thus, similar to separable spaces in Tversky and Gati (1982), violate the triangle inequality. One possibility is that dimensions were actually perceived integrally (Garner, 1974; Maddox, 1992; Melara, 1992). Importantly, our ideal observer simulations showed that satisfaction of ordinal triangle inequality does not necessitate satisfaction of segmental additivity. Thus, although our data leave open the possibility that such stimulus spaces are represented metrically, they do not rule out that they are not metric either, given the possibility of non-linear physical-to-psychological distance mapping.

In summary, evidence from behavioral similarity studies do not suggest that perceptual and conceptual stimuli are represented according to classical geometric theories. Among augmented geometric models, the predictions of the distance-density model have not panned out in several studies. Although the attention-weighted geometric model could explain some patterns in violations of the triangle inequality, the wider literature suggests

non-linearities between physical-to-psychological distance mapping, which violate the segmental additivity property of metric models, and which cannot be accounted for by attention-weighted models.

6.1.2 Challenging the validity of behavioral similarity tasks

The experiments in this thesis and previous studies that contested geometric models have relied on behavioural measures of similarity. One strategy to defend geometric models could be to argue that such behavioural readouts are either hopelessly noisy (per suggestions of Goodman (1972) and others, see Chapter 1 section 1.6 for discussion), or inherently do not reflect distances between mental representations. Perhaps psychological representations are actually geometric and do occupy some high-dimensional abstract space characterised by metric inter-item distances. However, once a behavioural readout of such distances is demanded, downstream psychological processes introduce inherent non-linearities that results in final data that do not adhere to metric principles.

First, such a defence would severely limit applicability and explanatory power of geometric models. To the extent that geometric theories cannot explain behaviour, but can only characterise mental representations, they lose the original allure and attraction that gained them so much popularity.

Second, one way around the noisiness of behavioural measures would be to directly compare neural representations of passively viewed objects. Techniques such as fMRI repetition suppression (Grill-Spector et al., 2006) or representational similarity analysis (RSA; Kriegeskorte, 2008) provide continuous measures of "distances" between representations, which would allow direct interval tests for geometric axioms, instead of having to rely on workarounds such as the ordinal triangle inequality test developed by Tversky and Gati. Such neural analysis methods were used by Theves et al. (2019) to calculate neural distances between items and to correlate them with their Euclidean distances in the underlying 2D stimulus space, arguing that the hippocampus provided a metric for cognitive spaces beyond physical space (also see Theves et al. 2020). However, they did not test whether such neural representations satisfied metric axioms. For example, symmetry could be tested by comparing the measures of repetition-suppression depending on the ordering of item presentations; segmental additivity could be measured using any three items on a straight line; while the triangle inequality could be assessed with any items forming a triangular arrangement in the stimulus space.

6.1.3 Representations of spatial and non-spatial schemas

The final two empirical chapters in this thesis take up the question of representation of more structured knowledge, such as spatial and non-spatial schemas. Chapter 4 develops a non-spatial schema generalisation paradigm, which could be further refined to systematically examine, for example, psychological processes underlying transfer from non-spatial to spatial knowledge. Chapter 5 presents two experiments that examine the nature of spatial schemas, introducing a learning task and experimental manipulations that could be easily adapted to study analogous questions within the non-spatial knowledge domain.

In Chapter 4, we reasoned that in a 2D stimulus space such as the bird space used by Constantinescu et al. (2016) and our experiments in Chapter 3, knowledge of associations between specific exemplars and some arbitrary "reward" stimuli could be viewed as "landmarks" in the stimulus space. Knowledge of the geometric shape of such landmark arrangements in one stimulus space could act as a non-spatial schema and could influence learning in a different stimulus space with similar landmark arrangements. However, across 2 experiments, we found that generalisation depended on the order of arrangements in our experimental sequence of conditions. Experiment 1 indicated that these order effects were likely due to certain boundary stimuli in one of the arrangements, which caused ceiling effects during learning. While Experiment 2 was designed to avoid such boundary stimuli, we still found an interaction between generalisation and order, with a possible explanation that some of the paired-associates were still too close to the corners of our 2D stimulus space. If the boundary stimuli are indeed the culprit, these effects might be related to the boundary effects found in our 1D stimulus space of Chapter 2. Given that our results and prior literature indicate that boundary stimuli are highly distinctive (Murdock, 1960), our recommendation for future experiments would be to employ a wider variation along the dimensions such as to steer far clear of boundaries.

With further refinement of our paradigm, establishment of a fast and efficient online generalisation task would contribute to answering several outstanding questions facing various cognitive disciplines. Bellmund and colleagues (2018) argued that one of the intriguing questions for the study of conceptual cognitive spaces is how trajectories encoded in one space can be retrieved to influence navigation in another space. To test the parallels between spatial and non-spatial reasoning, our non-spatial paired-associates learning task could be easily adapted for spatial stimulus-location learning. Then, one could explore the degree to which generalisation occurs across various geometric

operations on the spatial/non-spatial schemas in the two domains, such as rotation of the landmarks, or reflection across the axes. This would help establish whether schemas represent relationships between landmarks and/or relationships between landmarks and the boundaries of the spaces. Furthermore, these questions are highly relevant to the rich literature on analogical reasoning (Holyoak, 2012), where knowledge transfer across domains is thought to be guided by alignment of elements at multiple levels, including surface level perceptual and semantic similarities as well as deeper structural and functional similarities features (Gentner, 1983; Gentner & Markman, 1997; Holyoak & Koh, 1987). Precise characterisation of how agreement or mismatch of alignment across these different levels influences knowledge transfer is still an open challenge, whereby a rapid generalisation paradigm such as proposed in Chapter 4 could make a significant contribution. Finally, categorisation literature has suggested that within 2D psychologically separable stimulus spaces, when learning a categorisation decision boundary necessitates integration of information across dimensions, analogical transfer of such rules to a different part of the perceptual space is scant (Casale et al., 2012). However, if the categorisation rule is defined by a single dimension allowing verbalizable hypothesis-testing learning strategy, generalisation is seamless. Our paradigm could test if such results extend to analogical transfer *across* stimulus spaces, including transfer between spatial and non-spatial domains. Furthermore, using our design, future experiments could characterise how the nature of dimensions (i.e. psychologically separable versus integral) could interact with the relative contributions of implicit vs explicit learning systems and the subsequent success in generalisation (Ashby et al., 1998).

In Chapter 5, we examined the nature of representation of spatial schemas and their influence on learning within the spatial domain, presenting a paradigm that can be adapted for studying non-spatial schemas. Across 2 experiments, we found that stimulus-location associations on 2D boards have local facilitatory effect on learning of new such associations that are nearby, with no beneficial effects on learning of far-away paired-associates. Previous schema paradigms that have found drastic acceleration of learning argued that the source of this speed-up was an existing network of interconnected knowledge which acted as a unified schema (Guo & Yang, 2020, 2022; Sommer, 2017; Tse et al., 2007, 2011; van Buuren et al., 2014; Wang et al., 2012). Our results suggest, instead, that each individual element is learned separately without forming an

interconnected network or a "schema", and exerts local influence on learning of new information.

What is more, we found that even if stimulus-location associations regularly swap places, simply knowing which locations are designated for the stimuli is enough to exert local facilitatory effect on learning. Although this conflicts with previous studies that have found advantage for "consistent" conditions in which the object-locations pairs were fixed, relative to "inconsistent" conditions in which the locations were fixed but the objects rotated around them, we have argued that this likely stems from differences in our experimental designs. Unlike previous studies (e.g. Guo & Yang 2020), our paradigm did not have a separate stage for learning the initial paired-associates. Future studies should re-introduce such a stage, while adopting our distance manipulations between existing and to-be-learned paired-associates in order to further characterise differences between these conditions.

Finally, we propose that the learning paradigm and manipulations in Chapter 5 should be adapted for studying similar questions in non-spatial domains. Specifically, as in Constantinescu et al. (2016) and our experiments in Chapter 4, participants can be taught stimulus-exemplar associations in 2D stimulus spaces. Such non-spatial associative knowledge can subsequently be examined for their local vs global facilitatory influence on learning of novel paired-associates within the same stimulus space. Our prediction would be that, similar to spatial domain, non-spatial paired-associates will also act as independent "landmarks" in an abstract stimulus space, exerting facilitatory influence in their local neighbourhood.

6.1.4 Conclusion

The first two chapters of this thesis presented work that examined geometric theories of conceptual representation, which have experienced a recent resurgence due to discovered parallels between neural coding principles underlying physical spatial and non-spatial reasoning. We argue that conceptual knowledge is likely not represented as regions in a high-dimensional space with underlying metric organisational principles. This limits its parallels to representations of physical space, which have been argued to be Euclidean and supported by metric computations afforded by the grid cell system (Hafting et al., 2005; McNaughton et al., 2006; O'Keefe John, 1978). However, much of higher-level structured reasoning that occurs in physical versus conceptual domains likely still share similar psychological and neural computations, and findings in one domain should be

directly relevant to guide research in the other. In this spirit, the last two chapters presented paradigms that examine effects of spatial and non-spatial associative knowledge on learning. We hope that such experimental setups and manipulations can be further adapted to study similarities between representations of spatial and non-spatial knowledge.

7 REFERENCES

Aisbett, J., & Gibbon, G. (1994). A tunable distance measure for coloured solid models. *Artificial Intelligence*, *65*(1), 143–164. https://doi.org/10.1016/0004-3702(94)90039-6

Appelman, I. B., & Mayzner, M. S. (1982). Application of geometric models to letter recognition: Distance and density. *Journal of Experimental Psychology: General*, *111*(1). https://doi.org/10.1037/0096-3445.111.1.60

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481. https://doi.org/10.1037/0033-295X.105.3.442

Attneave, F. (1950). Dimensions of Similarity. *The American Journal of Psychology*, *63*(4), 516. https://doi.org/10.2307/1418869

Balkenius, C., & Gärdenfors, P. (2016). Spaces in the Brain: From Neurons to Meanings. *Frontiers in Psychology*, *7*(NOV), 1–12. https://doi.org/10.3389/fpsyg.2016.01820

Bao, X., Gjorgieva, E., Shanahan, L. K., Howard, J. D., Kahnt, T., & Gottfried, J. A. (2019). Grid-like Neural Representations Support Olfactory Navigation of a Two-Dimensional Odor Space. *Neuron*, *102*(5), 1066-1075.e5. https://doi.org/10.1016/j.neuron.2019.03.034

Bartlett, F. C. (1932). Remembering: A Study in Experimental and Social Psychology. Cambridge University Press.

Bassok, M., & Novick, L. R. (2012). Problem Solving. In *The Oxford Handbook of Thinking and Reasoning* (pp. 413–432). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.013.0021

Beals, R., & Krantz, D. H. (1967). Metrics and geodesics induced by order relations. *Mathematische Zeitschrift*, *101*(4), 285–298. https://doi.org/10.1007/BF01115107

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review*, *75*(2), 127–142. https://doi.org/10.1037/h0025470

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*(6415). https://doi.org/10.1126/science.aat6766

Blender Online Community. (2018). *Blender - a 3D modelling and rendering package*. Stichting Blender Foundation, Amsterdam. http://www.blender.org

Blumenthal, A., Duke, D., Bowles, B., Gilboa, A., Rosenbaum, R. S., Köhler, S., & McRae, K. (2017). Abnormal semantic knowledge in a case of developmental amnesia. *Neuropsychologia*, *102*, 237–247. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2017.06.018

Blumstein, S. E., & Stevens, K. N. (1981). Phonetic features and acoustic invariance in speech. *Cognition*, *10*(1–3), 25–32. https://doi.org/10.1016/0010-0277(81)90021-4

Bokeria, L., Henson, R. N., & Mok, R. M. (2021). Map-Like Representations of an Abstract Conceptual Space in the Human Brain. *Frontiers in Human Neuroscience*, *15*. https://doi.org/10.3389/fnhum.2021.620056

Borg, I., & Groenen, P. J. F. (2005). Modern multidimensional scaling: Theory and applications, 2nd ed. In *Modern multidimensional scaling: Theory and applications, 2nd ed.*

Bozeat, S., Lambon Ralph, M. A., Patterson, K., Garrard, P., & Hodges, J. R. (2000). Non-verbal semantic impairment in semantic dementia. *Neuropsychologia*, *38*. https://doi.org/10.1016/S0028-3932(00)00034-8

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4). https://doi.org/10.1163/156856897X00357

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) Phase II: 930 New Normative Photos. *PLoS ONE*, *9*(9), e106953. https://doi.org/10.1371/journal.pone.0106953

Burns, B., Shepp, B. E., McDonough, D., & Wiener-Ehrlich, W. K. (1978). The Relation Between Stimulus Analyzability and Perceived Dimensional Structure. *Psychology*

of Learning and Motivation - Advances in Research and Theory, *12*(C). https://doi.org/10.1016/S0079-7421(08)60008-0

Buzsáki, G., & Moser, E. I. (2013). Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature Neuroscience*, *16*(2), 130–138. https://doi.org/10.1038/nn.3304

Capitani, E., Laiacona, M., Mahon, B., & Caramazza, A. (2003). What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. *Cogn. Neuropsychol.*, *20*. https://doi.org/10.1080/02643290244000266

Carnap, R. (1928). *The logical structure of the world*. University of California Press.

Carroll, J. D., & Arabie, P. (1980). Multidimensional Scaling. *Annual Review of Psychology*, *31*(1), 607–649. https://doi.org/10.1146/annurev.ps.31.020180.003135

Carroll, J. D., & Wish, M. (1974). Multidimensional perceptual models and measurement methods. *Handbook of Perception*, *2*, 391–447.

Casale, M. B., Roeder, J. L., & Ashby, F. G. (2012). Analogical transfer in perceptual categorization. *Memory & Cognition*, *40*(3), 434–449. https://doi.org/10.3758/s13421-011-0154-4

Catrambone, R., Craig, D. L., & Nersessian, N. J. (2006). The role of perceptually represented structure in analogical problem solving. *Memory and Cognition*, *34*(5). https://doi.org/10.3758/BF03193258

Cheng, P. W., & Buehner, M. J. (2012). Causal learning. In *The Oxford handbook of thinking and reasoning.* (pp. 210–233). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.001.0001

Chouinard, P. A., & Goodale, M. A. (2009). Category-specific neural processing for naming pictures of animals and naming pictures of tools: an ALE meta-analysis. *Neuropsychologia*, *48*. https://doi.org/10.1016/j.neuropsychologia.2009.09.032

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., & Gabrieli, J. D. E. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *NeuroImage*, *14*(5). https://doi.org/10.1006/nimg.2001.0922

Cohen, N. J., & Eichenbaum, H. (1993). Memory, amnesia, and the hippocampal system. In *Memory, amnesia, and the hippocampal system.* The MIT Press.

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science*, *210*(4466). https://doi.org/10.1126/science.7414331

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Collins, A., & Quilliam, R. (1972). How to make a language user. In *Organization of memory*.

Collins, E., & Behrmann, M. (2020). Exemplar learning reveals the representational origins of expert category perception. *Proceedings of the National Academy of Sciences*, *117*(20). https://doi.org/10.1073/pnas.1912734117

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941

Coombs, C. H. (1954). A method for the study of interstimulus similarity. *Psychometrika*, *19*(3). https://doi.org/10.1007/BF02289183

Corter, J. E. (1987). Similarity, Confusability, and the Density Hypothesis. *Journal of Experimental Psychology: General*, *116*(3). https://doi.org/10.1037/0096-3445.116.3.238

Corter, J. E. (1988). Testing the Density Hypothesis: Reply to Krumhansl. *Journal of Experimental Psychology: General*, *117*(1). https://doi.org/10.1037/0096-3445.117.1.105

Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., & Damasio, A. R. (1996). A neural basis for lexical retrieval. *Nature*, *380*. https://doi.org/10.1038/380499a0

Davis, T., Love, B. C., & Preston, A. R. (2012a). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex*, *22*(2), 260–273. https://doi.org/10.1093/cercor/bhr036

Davis, T., Love, B. C., & Preston, A. R. (2012b). Striatal and hippocampal entropy and recognition signals in category learning: Simultaneous processes revealed by model-based fMRI. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(4). https://doi.org/10.1037/a0027865

Davis, T., Xue, G., Love, B. C., Preston, A. R., & Poldrack, R. A. (2014). Global neural pattern similarity as a common basis for categorization and recognition memory. *Journal of Neuroscience*, *34*(22). https://doi.org/10.1523/JNEUROSCI.3376-13.2014

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6). https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12). https://doi.org/10.1038/nn1560

Day, S. B., & Goldstone, R. L. (2011). Analogical Transfer From a Simulated Physical System. *Journal of Experimental Psychology: Learning Memory and Cognition*, *37*(3), 551–567. https://doi.org/10.1037/a0022333

Decock, L., & Douven, I. (2011). Similarity After Goodman. *Review of Philosophy and Psychology*, *2*(1), 61–75. https://doi.org/10.1007/s13164-010-0035-y

Demiralp, Ç., Bernstein, M. S., & Heer, J. (2014). Learning perceptual kernels for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12). https://doi.org/10.1109/TVCG.2014.2346978

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, *463*(7281), 657–661. https://doi.org/10.1038/nature08704

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179. https://doi.org/10.1016/j.tics.2010.01.004

Eichenbaum, H., & Cohen, N. J. (2004). *From Conditioning to Conscious Recollection*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195178043.001.0001

Eichenbaum, H., & Cohen, N. J. (2014). Can We Reconcile the Declarative Memory and Spatial Navigation Views on Hippocampal Function? In *Neuron* (Vol. 83, Issue 4, pp. 764–770). https://doi.org/10.1016/j.neuron.2014.07.032

Epstein, R. A., Patai, E. Z., Julian, J. B., & Spiers, H. J. (2017). The cognitive map in humans: spatial navigation and beyond. *Nature Neuroscience*, *20*(11), 1504–1513. https://doi.org/10.1038/nn.4656

Farzanfar, D., Spiers, H. J., Moscovitch, M., & Rosenbaum, R. S. (2022). From cognitive maps to spatial schemas. *Nature Reviews Neuroscience*. https://doi.org/10.1038/s41583-022-00655-9

Fechner, G. T. (1860). *Elemente der psychophysik* . Breitkopf und Härtel.

Fernández, G., & Morris, R. G. M. (2018). Memory, Novelty and Prior Knowledge. *Trends in Neurosciences*, *41*(10), 654–659. https://doi.org/10.1016/j.tins.2018.08.006

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*(2). https://doi.org/10.1016/0364-0213(95)90016-0

Gärdenfors, P. (2000). *Conceptual Spaces*. The MIT Press. https://doi.org/10.7551/mitpress/2076.001.0001

Garner, W. R. (1974). The Processing of Information and Structure. In *The Processing of Information and Structure*. Psychology Press. https://doi.org/10.4324/9781315802862

Garner, W. R. (1976). Interaction of stimulus dimensions in concept and choice processes. *Cognitive Psychology*, *8*(1). https://doi.org/10.1016/0010-0285(76)90006-2

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *ELife*, *6*, 1–20. https://doi.org/10.7554/eLife.17086

Gati, I., & Tversky, A. (1982). Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(2), 325–340. https://doi.org/10.1037/0096-1523.8.2.325

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170. https://doi.org/10.1016/S0364-0213(83)80009-3

Gentner, D., & Markman, A. B. (1997). Structural alignment in analogy and similarity. *American Psychologist*, *52*(52), 45–56.

https://pdfs.semanticscholar.org/7d65/5b765d962d47fa3eeb677c7411056b38165b.pdf

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The Roles of Similarity in Transfer: Separating Retrievability From Inferential Soundness. *Cognitive Psychology*, *25*(4). https://doi.org/10.1006/cogp.1993.1013

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*(1), 104–114. https://doi.org/10.1016/j.neuropsychologia.2013.11.010

Gibson, E. J. (1969). Principles of Perceptual Learning and Development. Appleton-Century-Crofts.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306–355. https://doi.org/10.1016/0010-0285(80)90013-4

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analagocial transfer. *Cognitive Psychology*, *15*.

Gilboa, A., & Marlatte, H. (2017). Neurobiology of Schemas and Schema-Mediated Memory. *Trends in Cognitive Sciences*, *21*(8), 618–631. https://doi.org/10.1016/j.tics.2017.04.013

Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics*, *25*(5), 425–431. https://doi.org/10.3758/BF03199852

Goldstone, R. L., Medin, D. L., & Halberstadt, J. (1997). Similarity in context. *Memory & Cognition*, *25*(2), 237–255. https://doi.org/10.3758/BF03201115

Goldstone, R. L., & Son, J. Y. (2012). Similarity. In *The Oxford Handbook of Thinking and Reasoning* (pp. 155–176). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.013.0010

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects*. Bobbs-Merrill.

Green, A. E., Fugelsang, J. A., Kraemer, D. J. M., Shamosh, N. A., & Dunbar, K. N. (2006). Frontopolar cortex mediates abstract integration in analogy. *Brain Research*, *1096*(1), 125–137. https://doi.org/10.1016/j.brainres.2006.04.024

Green, A. E., Kraemer, D. J. M., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex*, *20*(1). https://doi.org/10.1093/cercor/bhp081

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23. https://doi.org/10.1016/j.tics.2005.11.006

Guido, V. R., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Guo, D., & Yang, J. (2020). Interplay of the long axis of the hippocampus and ventromedial prefrontal cortex in schema-related memory retrieval. *Hippocampus*, *30*(3). https://doi.org/10.1002/hipo.23154

Guo, D., & Yang, J. (2022). Reactivation of schema representation in lateral occipital cortex supports successful memory encoding. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhac475

Hafting, T., Fyhn, M., Molden, S., Moser, M. B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052). https://doi.org/10.1038/nature03721

Hardiman, P. T., Dufresne, R., & Mestre, J. P. (1989). The relation between problem categorization and problem solving among experts and novices. *Memory & Cognition*, *17*(5). https://doi.org/10.3758/BF03197085

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, *56*(1). https://doi.org/10.1037/h0062474

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11). https://doi.org/10.1038/s41562-020-00951-3

Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior*, *8*(2). https://doi.org/10.1016/S0022-5371(69)80058-7

Hodges, J. R., & Patterson, K. (2007). Semantic dementia: a unique clinicopathological syndrome. *Lancet Neurol.*, *6*. https://doi.org/10.1016/S1474-4422(07)70266-1

Holland, J. H., Holyoak, K. J., & Nisbett, R. E. (1986). *Induction: Processes of Inference, Learning, and Discovery*. MIT Press.

Holman, E. W. (1979). Monotonic models for asymmetric proximities. *Journal of Mathematical Psychology*, *20*. https://doi.org/10.1016/0022-2496(79)90031-2

Holyoak, K. J. (2012). Analogy and Relational Reasoning. In *The Oxford Handbook of Thinking and Reasoning*. https://doi.org/10.1093/oxfordhb/9780199734689.013.0013

Holyoak, K. J., & Cheng, P. W. (2011). Causal Learning and Inference as a Rational Process: The New Synthesis. *Annual Review of Psychology*, *62*(1), 135–163. https://doi.org/10.1146/annurev.psych.121208.131634

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15*(4), 332–340. https://doi.org/10.3758/BF03197035

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*(3). https://doi.org/10.1016/0364-0213(89)90016-5

Hosman, J., & Künnapas, T. (1972). *On the Relation Between Similarity and Dissimilarity Estimates*. Psychological Laboratories, University of Stockholm.

Houston, R. A., & Shearer, J. F. (1930). Fechner's Law. *Nature*, *125*(3163), 891–892. https://doi.org/10.1038/125891b0

Hummel, J. E., & Holyoak, K. J. (1997). Distributed Representations of Structure: A Theory of Analogical Access and Mapping. *Psychological Review*, *104*(3). https://doi.org/10.1037/0033-295X.104.3.427

Hutchinson, J. W., & Lockhead, G. R. (1977). Similarity as distance: A structural principle for semantic memory. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(6). https://doi.org/10.1037/0278-7393.3.6.660

Jacobs, J., Weidemann, C. T., Miller, J. F., Solway, A., Burke, J. F., Wei, X.-X., Suthana, N., Sperling, M. R., Sharan, A. D., Fried, I., & Kahana, M. J. (2013). Direct recordings of grid-like neuronal activity in human spatial navigation. *Nature Neuroscience*, *16*(9), 1188–1190. https://doi.org/10.1038/nn.3466

Jakobson, R., Fant, G., & Halle, M. (1961). *Preliminaries to Speech Analysis*. The MIT Press.

James, W. (1890). The Principles Of Psychology Volume I By William James (1890). *The Principles of Psychology*, *I*(1890).

Jefferies, E., Patterson, K., Jones, R. W., & Lambon Ralph, M. A. (2009). Comprehension of concrete and abstract words in semantic dementia. *Neuropsychology*, *23*. https://doi.org/10.1037/a0015452

Jeffreys, H. (1998). *The Theory of Probability*. Oxford University Press.

Jones, M., Maddox, W., & Love, B. (2006). The role of similarity in generalization. *Proceedings of the 28th Annual Meeting of the Cognitive Society*.

Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. *Proc. Natl Acad. Sci. USA*, *107*. https://doi.org/10.1073/pnas.1005062107

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773. https://doi.org/10.2307/2291091

Keren, G., & Baggen, S. (1981). Recognition models of alphanumeric characters. *Perception & Psychophysics*, *29*(3), 234–246. https://doi.org/10.3758/BF03207290

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3. *Perception 36 ECVP Abstract Supplement*.

Knierim, J. J., Neunuebel, J. P., & Deshmukh, S. S. (2014). Functional correlates of the lateral and medial entorhinal cortex: objects, path integration and local–global reference frames. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1635), 20130369. https://doi.org/10.1098/rstb.2013.0369

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, *262*(5140), 1747–1749. https://doi.org/10.1126/science.8259522

Knudsen, E. B., & Wallis, J. D. (2021). Hippocampal neurons construct a map of an abstract value space. *Cell*, *184*(18). https://doi.org/10.1016/j.cell.2021.07.010

Komorowski, R. W., Manns, J. R., & Eichenbaum, H. (2009). Robust conjunctive item - Place coding by hippocampal neurons parallels learning what happens where. *Journal of Neuroscience*, *29*(31). https://doi.org/10.1523/JNEUROSCI.1378-09.2009

Krantz, D. H., & Tversky, A. (1975). Similarity of rectangles: An analysis of subjective dimensions. *Journal of Mathematical Psychology*, *12*(1), 4–34. https://doi.org/10.1016/0022-2496(75)90047-4

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(NOV). https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, *3*(JUL). https://doi.org/10.3389/fpsyg.2012.00245

Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, *85*(5), 445–463. https://doi.org/10.1037/0033-295X.85.5.445

Krumhansl, C. L. (1988). Testing the Density Hypothesis: Comment on Corter. *Journal of Experimental Psychology: General*, *117*(1). https://doi.org/10.1037/0096-3445.117.1.101

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1). https://doi.org/10.1007/BF02289565

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2). https://doi.org/10.1007/BF02289694

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the Emergence of Conceptual Knowledge during Human Decision Making. *Neuron*, *63*(6), 889–901. https://doi.org/10.1016/J.NEURON.2009.07.030

Lambon Ralph, M. A. (2014). Neurocognitive insights on conceptual knowledge and its breakdown. *Phil. Trans. R. Soc. B*, *369*. https://doi.org/10.1098/rstb.2012.0392

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). Generalisation and differentiation in semantic memory: insights from semantic dementia. *Ann. NY Acad. Sci.*, *1124*. https://doi.org/10.1196/annals.1440.006

Lambon Ralph, M. A., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proc. Natl Acad. Sci. USA*, *107*. https://doi.org/10.1073/pnas.0907307107

Li, L., Malave, V., Song, A., & Yu, A. J. (2016). Extracting Human Face Similarity Judgments: Pairs or Triplets? *Proceedings of the 38th Annual Meeting of the Cognitive Science Society, CogSci 2016*. https://doi.org/10.1167/16.12.719

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

Love, B. C., & Roads, B. D. (2021). Similarity as a Window on the Dimensions of Object Representation. In *Trends in Cognitive Sciences* (Vol. 25, Issue 2). https://doi.org/10.1016/j.tics.2020.12.003

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208. https://doi.org/10.1073/pnas.1614048113

Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters*, *680*, 31–38. https://doi.org/10.1016/j.neulet.2017.07.061

Maddox, W. T. (1992). Perceptual and decisional separability. In *Multidimensional models of perception and cognition.* (pp. 147–180). Lawrence Erlbaum Associates, Inc.

Mahon, B. Z., Anzellotti, S., Schwarzbach, J., Zampini, M., & Caramazza, A. (2009). Category-specific organization in the human brain does not require visual experience. *Neuron*, *63*. https://doi.org/10.1016/j.neuron.2009.07.012

Manns, J. R., & Eichenbaum, H. (2006). Evolution of declarative memory. *Hippocampus*, *16*(9), 795–808. https://doi.org/10.1002/hipo.20205

Markman, A. B. (2012). Knowledge Representation. In *The Oxford Handbook of Thinking and Reasoning* (pp. 36–51). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.013.0004

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. MIT Press.

Mathy, F., Haladjian, H. H., Laurent, E., & Goldstone, R. L. (2013). Similarity-dissimilarity competition in disjunctive classification tasks. *Frontiers in Psychology*, *4*(FEB). https://doi.org/10.3389/fpsyg.2013.00026

MATLAB. (2020). *version R2020a.* Natick, Massachusetts: The MathWorks Inc.

McClelland, J. L. (2013). Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General*, *142*(4), 1190–1210. https://doi.org/10.1037/a0033812

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I., & Moser, M.-B. (2006). Path integration and the neural basis of the 'cognitive map'. *Nature Reviews Neuroscience*, *7*(8), 663–678. https://doi.org/10.1038/nrn1932

Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science*, *1*(1). https://doi.org/10.1111/j.1467-9280.1990.tb00069.x

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254–278. https://doi.org/10.1037/0033-295X.100.2.254

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3). https://doi.org/10.1037/0033-295X.85.3.207

Melara, R. D. (1992). The Concept of Perceptual Similarity: From Psychophysics to Cognitive Psychology. *Advances in Psychology*, *92*(C). https://doi.org/10.1016/S0166-4115(08)61782-3

Mok, R. M., & Love, B. C. (2019). A non-spatial account of place and grid cells based on clustering models of concept learning. *Nature Communications*, *10*(1), 5685. https://doi.org/10.1038/s41467-019-13760-8

Monahan, J. S., & Lockhead, G. R. (1977). Identification of integral stimuli. *Journal of Experimental Psychology: General*, *106*(1). https://doi.org/10.1037/0096-3445.106.1.94

Morey, R. D., & Rouder, J. N. (2021). *BayesFactor: Computation of Bayes Factors for Common Designs*. https://CRAN.R-project.org/package=BayesFactor

Morton, N. W., & Preston, A. R. (2021). Concept formation as a computational cognitive process. In *Current Opinion in Behavioral Sciences* (Vol. 38). https://doi.org/10.1016/j.cobeha.2020.12.005

Moser, E. I., Moser, M.-B., & McNaughton, B. L. (2017). Spatial representation in the hippocampal formation: a history. *Nature Neuroscience*, *20*(11), 1448–1464. https://doi.org/10.1038/nn.4653

Murdock, B. B. (1960). The distinctiveness of stimuli. *Psychological Review*, *67*(1). https://doi.org/10.1037/h0042382

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316. https://doi.org/10.1037/0033-295X.92.3.289

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10). https://doi.org/10.1038/s41593-019-0470-8

Nosofsky, R. M. (1985a). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics*, *38*(5). https://doi.org/10.3758/BF03207172

Nosofsky, R. M. (1985b). Luce's choice model and Thurstone's categorical judgment model compared: Kornbrot's data revisited. *Perception & Psychophysics*, *37*(1), 89–91. https://doi.org/10.3758/BF03207144

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. 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). https://doi.org/10.1037/0278-7393.13.1.87

Nosofsky, R. M. (1992a). Exemplar, prototypes, and similarity rules. From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes, Vol. 1.

Nosofsky, R. M. (1992b). Similarity Scaling and Cognitive Process Models. *Annual Review of Psychology*, *43*(1), 25–53. https://doi.org/10.1146/annurev.ps.43.020192.000325

O'Keefe John, N. L. (1978). *The Hippocampus as a Cognitive Map*. Clarendon Press.

Osgood, C. E. (1949). The similarity paradox in human learning: a resolution. *Psychological Review*, *56*(3). https://doi.org/10.1037/h0057488

Park, S. A., Miller, D. S., Nili, H., Ranganath, C., & Boorman, E. D. (2020). Map Making: Constructing, Combining, and Inferring on Abstract Cognitive Maps. *Neuron*, *107*(6), 1226-1238.e8. https://doi.org/10.1016/j.neuron.2020.06.030

Peer, M., Brunec, I. K., Newcombe, N. S., & Epstein, R. A. (2021). Structuring Knowledge with Cognitive Maps and Cognitive Graphs. In *Trends in Cognitive Sciences* (Vol. 25, Issue 1). https://doi.org/10.1016/j.tics.2020.10.004

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4). https://doi.org/10.1163/156856897X00366

Piaget, J. (1926). Language and Thought of the Child. In *Language and Thought of the Child*. Kegan Paul, Trench & Trubner.

Piaget, J. (1952). The Origins of Intelligence in Children. W.W. Norton & Co.

Polk, T. A., Behensky, C., Gonzalez, R., & Smith, E. E. (2002). Rating the similarity of simple perceptual stimuli: asymmetries induced by manipulating exposure frequency. *Cognition*, *82*(3), B75–B88. https://doi.org/10.1016/S0010-0277(01)00151-2

Quent, A. J. (2021). *Novelty, Prediction Error and Memory Encoding: Limitations of the Pimms Framework* [Doctoral thesis]. https://doi.org/https://doi.org/10.17863/CAM.78604

Quine, O. W. v. (1969). *Ontological relativity and other essays*. Columbia University Press.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raven, J. C. (1938). Progressive matrices: A perceptual test of intelligence. H.K. Lewis.

Raybaut, P. (2009). Spyder-documentation. *Available Online at: Pythonhosted. Org*.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, *12*(1), 1–20. https://doi.org/10.1016/S0022-5371(73)80056-8

Roads, B. D., & Love, B. C. (2021). Enriching ImageNet with Human Similarity Judgments and Psychological Embeddings. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR46437.2021.00355

Roads, B. D., & Mozer, M. C. (2019). Obtaining psychological embeddings through joint kernel and metric learning. *Behavior Research Methods*, *51*(5). https://doi.org/10.3758/s13428-019-01285-3

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist*, *45*(9). https://doi.org/10.1037/0003-066X.45.9.1043

Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition: a Parallel Distributed Processing Approach*. https://doi.org/10.7551/mitpress/6161.001.0001

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, *7*(4), 532–547. https://doi.org/10.1016/0010-0285(75)90021-3

Ross, B. H. (1989). Distinguishing Types of Superficial Similarities: Different Effects on the Access and Use of Earlier Problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(3). https://doi.org/10.1037/0278-7393.15.3.456

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, *53*(2), 94–101. https://doi.org/10.1037/h0041867

RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Rueckemann, J. W., Sosa, M., Giocomo, L. M., & Buffalo, E. A. (2021). The grid code for ordered experience. *Nature Reviews Neuroscience*, *22*(10), 637–649. https://doi.org/10.1038/s41583-021-00499-9

Schapiro, A. C., Kustner, L. v., & Turk-Browne, N. B. (2012). Shaping of Object Representations in the Human Medial Temporal Lobe Based on Temporal Regularities. *Current Biology*, *22*(17), 1622–1627. https://doi.org/10.1016/J.CUB.2012.06.056

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486–492. https://doi.org/10.1038/nn.3331

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schott, B. H., Wüstenberg, T., Lücke, E., Pohl, I.-M., Richter, A., Seidenbecher, C. I., Pollmann, S., Kizilirmak, J. M., & Richardson-Klavehn, A. (2019). Gradual acquisition of visuospatial associative memory representations via the dorsal precuneus. *Human Brain Mapping*, *40*(5), 1554–1570. https://doi.org/10.1002/hbm.24467

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron*, *91*(6), 1402–1412. https://doi.org/10.1016/j.neuron.2016.08.019

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4). https://doi.org/10.1007/BF02288967

Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, *55*(6), 509–523. https://doi.org/10.1037/h0042354

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140. https://doi.org/10.1007/BF02289630

Shepard, R. N. (1963). Analysis of Proximities as a Technique for the Study of Information Processing in Man. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *5*(1). https://doi.org/10.1177/001872086300500104

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*(1). https://doi.org/10.1016/0022-2496(64)90017-3

Shepard, R. N. (1980). Multidimensional Scaling, Tree-Fitting, and Clustering. *Science*, *210*(4468), 390–398. https://doi.org/10.1126/science.210.4468.390

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820). https://doi.org/10.1126/science.3629243

Sjöberg, L. (1972). A Cognitive Theory of Similarity. *Göteborg Psychological Reports*, *2*(10).

Sloman, S. A. (1993). Feature-Based Induction. *Cognitive Psychology*, *25*(2), 231–280. https://doi.org/10.1006/cogp.1993.1006

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*(3). https://doi.org/10.1037/h0036351

Smith, L. B., & Heise, D. (1992). Perceptual Similarity and Conceptual Structure. *Advances in Psychology*, *93*(C). https://doi.org/10.1016/S0166-4115(08)61009-2

Snowden, J. S., Goulding, P. J., & Neary, D. (1989). Semantic dementia: a form of circumscribed cerebral atrophy. *Behav. Neurol.*, *2*.

Sommer, T. (2017). The emergence of knowledge and how it supports the memory for novel related information. *Cerebral Cortex*, *27*(3). https://doi.org/10.1093/cercor/bhw031

Squire, L. R., & Knowlton, B. J. (1995). Learning about categories in the absence of memory. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(26). https://doi.org/10.1073/pnas.92.26.12470

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643–1653. https://doi.org/10.1038/nn.4650

Stone, J. v. (2013). Bayes' Rule: A tutorial introduction to Bayesian analysis. Sebtel Press.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*. MIT Press.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. T. (2011). Adaptively learning the crowd kernel. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*.

Tavares, R. M., Mendelsohn, A., Grossman, Y., Williams, C. H., Shapiro, M., Trope, Y., & Schiller, D. (2015). A Map for Social Navigation in the Human Brain. *Neuron*, *87*(1), 231–243. https://doi.org/10.1016/j.neuron.2015.06.011

Taylor, J. E., Cortese, A., Barron, H. C., Pan, X., Sakagami, M., & Zeithamova, D. (2021). *How do we generalize?* http://arxiv.org/abs/2104.00899

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. *Advances in Neural Information Processing Systems*.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640. https://doi.org/10.1017/S0140525X01000061

Thagard, P., Holyoak, K. J., Nelson, G., & Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence*, *46*(3). https://doi.org/10.1016/0004-3702(90)90018-U

Theves, S., Fernandez, G., & Doeller, C. F. (2019). The Hippocampus Encodes Distances in Multidimensional Feature Space. *Current Biology*, *29*(7), 1226-1231.e3. https://doi.org/10.1016/j.cub.2019.02.035

Theves, S., Fernández, G., & Doeller, C. F. (2020). The hippocampus maps concept space, not feature space. *Journal of Neuroscience*, *40*(38), 7318–7325. https://doi.org/10.1523/JNEUROSCI.0494-20.2020

Thorndike, E. L. (1931). *Human learning*. Century.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, *17*(4). https://doi.org/10.1007/BF02288916

Torgerson, W. S. (1965). Multidimensional scaling of similarity. *Psychometrika*, *30*(4), 379–393. https://doi.org/10.1007/BF02289530

Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, *9*(1), 40–50. https://doi.org/10.3758/BF03213026

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P., & Morris, R. G. M. (2007). Schemas and memory consolidation. *Science*, *316*(5821), 76–82. https://doi.org/10.1126/science.1135935

Tse, D., Takeuchi, T., Kakeyama, M., Kajii, Y., Okuno, H., Tohyama, C., Bito, H., & Morris, R. G. M. (2011). Schema-dependent gene activation and memory encoding in neocortex. *Science*, *333*(6044), 891–895. https://doi.org/10.1126/science.1205274

Tulving, E. (1972). Episodic and semantic memory. In *Organization of memory.* (pp. 423, xiii, 423–xiii). Academic Press.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*(4), 281–299. https://doi.org/10.1037/h0032955

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, *89*(2), 123–154. https://doi.org/10.1037/0033-295X.89.2.123

Tversky, A., & Krantz, D. H. (1969). Similarity of schematic faces: A test of interdimensional additivity. *Perception & Psychophysics*, *5*(2).

Tversky, A., & Krantz, D. H. (1970). The dimensional representation and the metric structure of similarity data. *Journal of Mathematical Psychology*, *7*(3). https://doi.org/10.1016/0022-2496(70)90041-6

van Buuren, M., Kroes, M. C. W., Wagner, I. C., Genzel, L., Morris, R. G. M., & Fernández, G. (2014). Initial investigation of the effects of an experimentally learned schema on spatial associative memory in humans. *Journal of Neuroscience*, *34*(50). https://doi.org/10.1523/JNEUROSCI.2365-14.2014

van Kesteren, M. T. R., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., & Fernández, G. (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: From congruent to incongruent. *Neuropsychologia*, *51*(12), 2352–2359. https://doi.org/10.1016/j.neuropsychologia.2013.05.027

van Kesteren, M. T. R., Fernández, G., Norris, D. G., & Hermans, E. J. (2010). Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(16), 7550–7555. https://doi.org/10.1073/pnas.0914892107

van Kesteren, M. T. R., Rignanese, P., Gianferrara, P. G., Krabbendam, L., & Meeter, M. (2020). Congruency and reactivation aid memory integration through reinstatement of prior knowledge. *Scientific Reports*, *10*(1), 4776. https://doi.org/10.1038/s41598-020-61737-1

van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211–219. https://doi.org/10.1016/j.tins.2012.02.001

Viganò, S., & Piazza, M. (2020). Distance and Direction Codes Underlie Navigation of a Novel Semantic Space in the Human Brain. *The Journal of Neuroscience*, *40*(13), 2727–2736. https://doi.org/10.1523/JNEUROSCI.1849-19.2020

Wang, S. H., Tse, D., & Morris, R. G. M. (2012). Anterior cingulate cortex in schema assimilation and expression. *Learning and Memory*, *19*(8), 315–318. https://doi.org/10.1101/lm.026336.112

Weber, E. H. (1851). Der Tastsinn und das Gemeingefühl.

Wender, K. (1971). A test of independence of dimensions in multidimensional scaling. *Perception & Psychophysics*, *10*(1). https://doi.org/10.3758/BF03205762

Wiener-Ehrlich, W. K. (1978). Dimensional and metric structures in multidimensional stimuli. *Perception & Psychophysics*, *24*(5), 399–414. https://doi.org/10.3758/BF03199737

Wikenheiser, A. M., & Schoenbaum, G. (2016). Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nature Reviews Neuroscience*, *17*(8), 513–523. https://doi.org/10.1038/nrn.2016.56

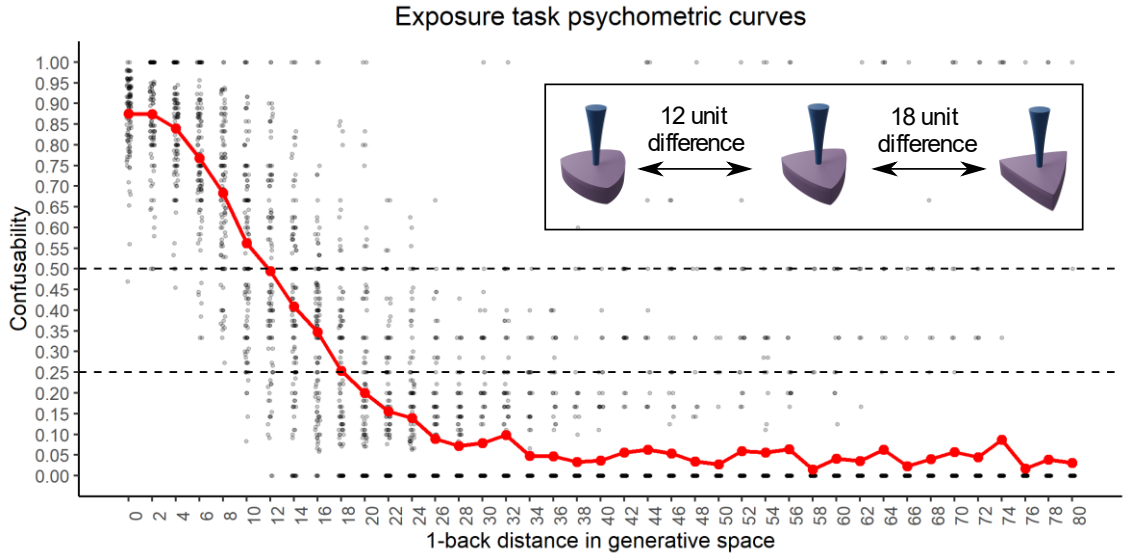
Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal Cortex as a Cognitive Map of Task Space. *Neuron*, *81*(2), 267–279. https://doi.org/10.1016/j.neuron.2013.11.005

8 APPENDICES

APPENDIX FOR CHAPTER 2

8.1 The psychometric curve for the same-different exposure task

Supplementary Supplementary Figure 8.1 shows the psychometric curve for the same-different task, with the confusability on the *y* axis measuring how often the participants responded "same" on trials with a particular 1-back distance between the exemplars. Due to lack of trials, we could not compare item-specific psychometric curves, to test whether items in high-density regions had steeper drop-off.



Supplementary Figure 8.1: Psychometric curves for the same-different exposure task.

Each dot is a participant's average confusability for two stimuli at a certain distance from each other in the generative space. Red dots indicate averages across participants. Dashed black lines denote 50% and 75% accuracy lines for reference. The inset image shows example unit differences at which the participants reach 50% or 75% accuracy.

The curve shows that the participants had difficulty differentiating close-by stimuli, with a 50% accuracy when the items were 12 units apart in the generative space, reaching 75% accuracy with 18 unit difference.

Perceptual discrimination is arguably harder during a 1-back task versus the triplet matching task when all the relevant stimuli are displayed on the screen at the same time.

Nevertheless, these results indicate that the participants probably also struggled to tell apart neighbouring stimuli in the generative space when doing the triplet task. Thus, although our experimental design aimed to only influence neighbourhood density and not the repetition frequency of each exemplar, the participants may have perceived certain neighbours as the same stimuli, producing a similar psychological effect as from the repetition of the same stimuli (and thus possibly influencing the saliency of stimuli, as in Polk et al. 2002). Future studies should ensure adequate discriminability between neighbourhood exemplars, to avoid possibly confounding stimulus salience increases from neighbourhood density increases.

APPENDIX FOR CHAPTER 3

8.2 The two-dimensional monotone proximity structure and its elementary principles

Tversky and Gati (1982) discussed various elementary properties that a two-dimensional structure (such as the 2D spaces used in their experiments) must satisfy. For two dimensions A and P, let A = {a,b,c,…} and P = {p,q,r,…} denote attributes, with A x P denoting the product set consisting of all combinations of ap, bq, bp, etc. (see Figure 3.1-B). Let δ(ap,bq) denote an ordinal dissimilarity measure or a psychological distance between the points. For such a two-dimensional proximity structure (A X P, δ) to be a metric representation, it must satisfy three elementary properties:

- Dominance: Two-dimensional difference exceeds each of the one-dimensional components. δ (ap,bq) > δ(ap,aq), δ(aq,bq).
- Consistency: Ordering of intervals on one attribute is independent of the fixed level of the other attribute.
- δ(ap,bp) > δ(cp,dp) iff δ(aq,bq) > δ(cq,dq)

  and

  δ(ap,aq) > δ(ar,as) iff δ(bp,bq) > δ(br,bs).
- Transitivity: Relation of betweenness is transitive or noncircular. If a|b|c denotes that b lies between a and c, then a|b|c and b|c|d imply a|b|d and a|c|d.

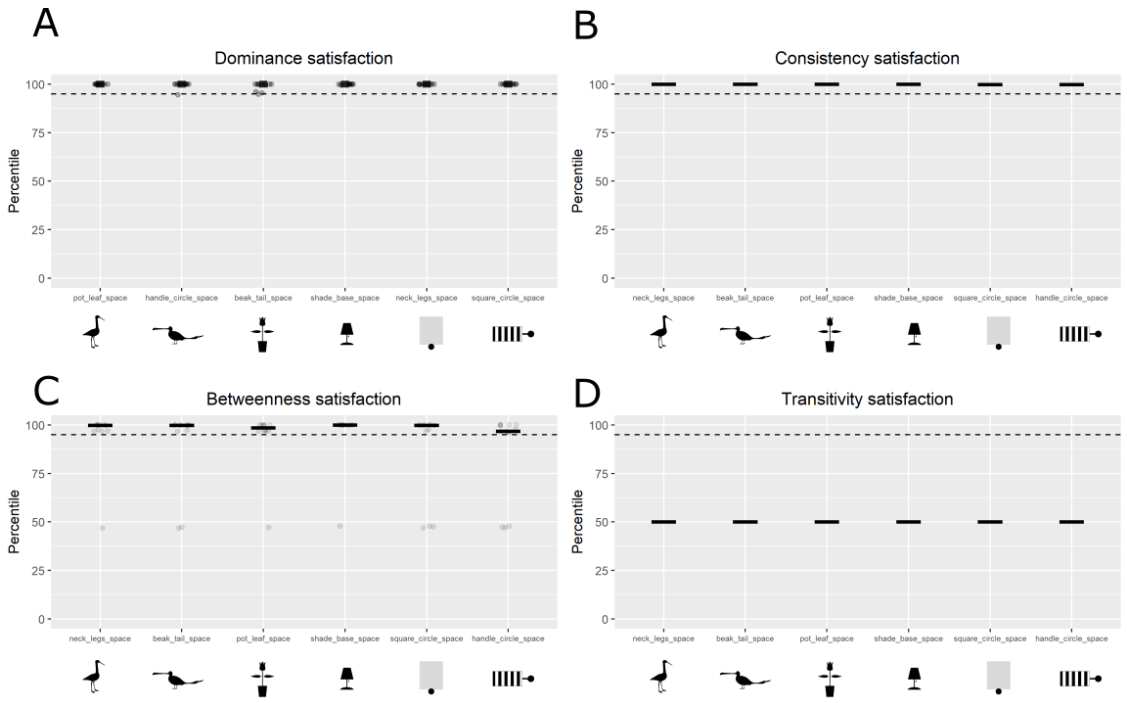
In their paper, Tversky and Gati (1982) first checked whether the data from the six studies they discuss satisfied the elementary properties above, before testing them for the triangle inequality satisfactions. They report that "The average dissimilarities from all six studies satisfied all the defining properties of a two-dimensional monotone proximity structure: dominance (Equation 1), consistency (Equation 2), and transitivity (Equation 3). The only violations occurred in Study 6 and they were relatively minor" (Tversky and Gati, 1982, p.130) It is unclear how the data were averaged, and how the satisfaction was tested statistically.

In our study, we tested the reported similarity values for the three elementary conditions as well. As for the main analyses reported above, for each participant, we first averaged data from repetitions 2 and 3.

For dominance, we compared each two-dimensional distance to its unidimensional components. For each participant, we counted the number of dominance satisfactions and compared it to the participant-specific null distribution derived from randomly permuting responses 10,000 times. Supplementary Figure 8.2 below shows the percentile value of each participant in each of our six stimulus groups. We can see all groups significantly satisfied the dominance condition.

For consistency, we checked for consistency satisfaction along both dimensions for each participant by counting the number of dimensions that satisfied consistency and comparing it to participant-specific null distribution. Supplementary Figure 8.2-B shows that consistency was satisfied for all groups.

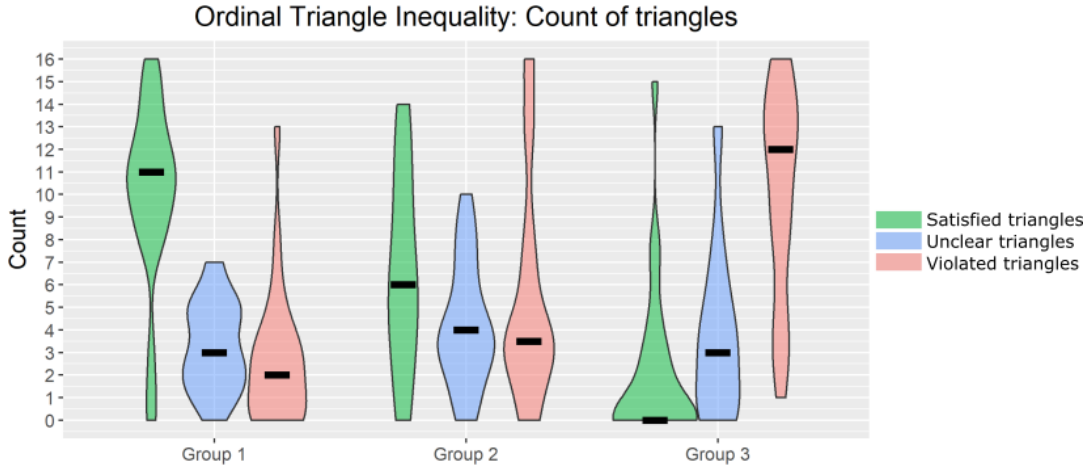
For transitivity, we first checked whether our participants satisfied the betweenness property. Supplementary Figure 8.2-C shows that the participants in all groups significantly satisfied the betweenness condition. However, the groups did not satisfy the transitivity property. Since violation of transitivity means the betweenness on smaller distances (for example, ap|bp|cp and bp|cp|dp) did not imply betweenness on larger distances (i.e. ap|bp|dp and ap|cp|dp), in all of our analysis we did not use data from judgments of pairs that were more than two levels apart in the space.



Supplementary Figure 8.2: Satisfaction of elementary properties for the 2D monotone proximity structure.

(A) Dominance. (B) Consistency. (C) Betweenness. (D) Transitivity. Black bars indicate median percentile values for each group. Dots represent participants. Dashed horizontal lines represent 95th percentile value for reference.

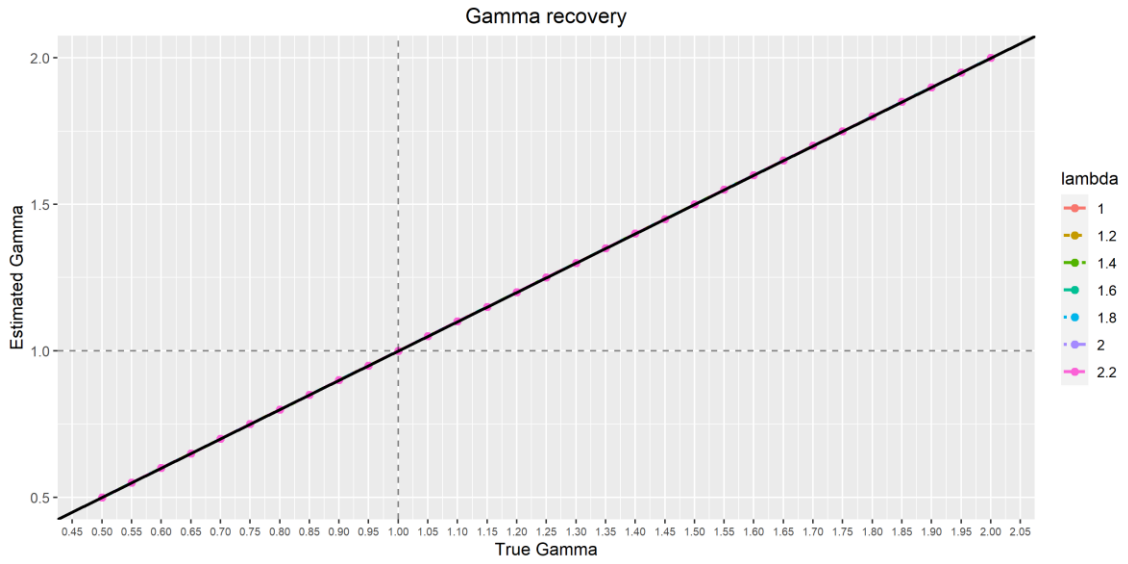
8.3 Count of satisfied, violated, or non-diagnostic triangles for the ordinal triangle inequality test



Supplemental Figure 8.3: Ordinal triangle inequality test outcomes.

Count of triangles satisfying, violating, or not providing sufficient information for checking ordinal triangle inequality. Black bars represent median values per group per condition.

8.4 Gamma recovery for continuous psychological distances

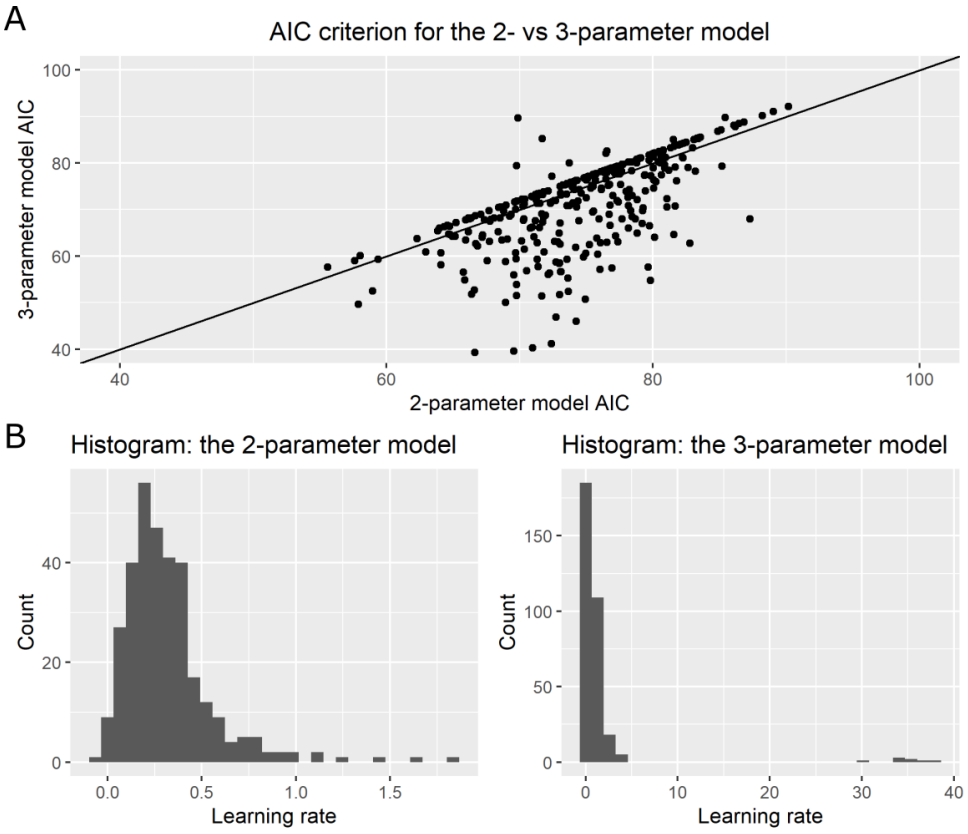


Supplemental Figure 8.4: γ estimation on continuous perceived dissimilarity values p_δ of ideal observers.
For all simulations, our procedure accurately recovered the true γ value.

APPENDIX FOR CHAPTER 5

8.5 Experiment 1: Comparison of 2-parameter and 3-parameter models

Each model was fit separately to the averaged PA data of each learning condition for each participant. Analysis of the AIC criterion showed that for the majority of participants and conditions (171 vs 154), the 3-parameter model outperformed the 2-parameter model (Supplementary Figure 8.5-A). However, estimates from the 3-parameter model were noisier as shown by large outlier values in the histogram distribution of learning rate estimates (Supplementary Figure 8.5-B). Therefore, we decided to go with the 2-parameter model for our data analysis[9].

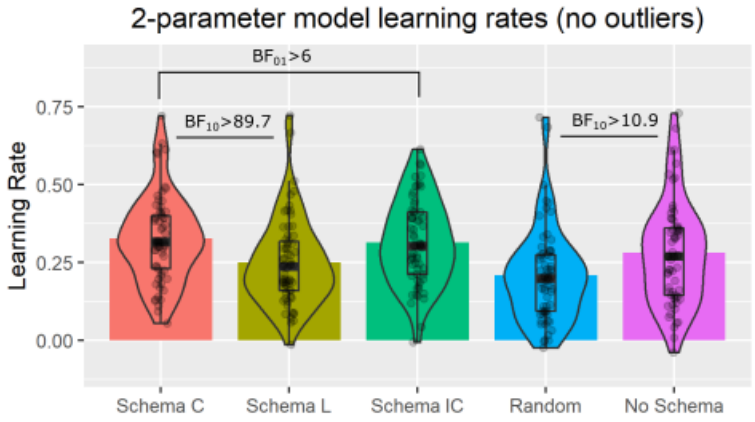


Supplementary Figure 8.5: Comparison of the 2-parameter and 3-parameter models.

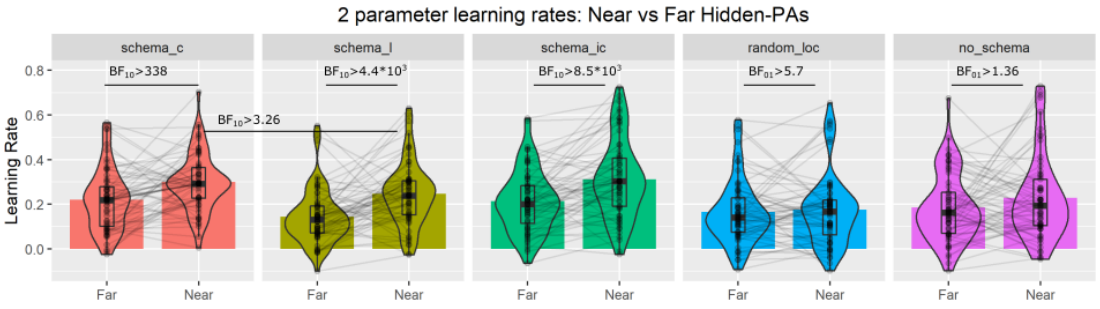
[9] A deviation from our pre-registration.

(A) AIC value scatterplots for 2 vs 3 parameter models. Each dot is a participant. Black line is a 45 degree line. (B) Histogram of learning rate *c* estimates for the 2-parameter model (Left) and the 3-parameter model (Right).

8.6 Experiment 1: 2-parameter model estimates for learning rates

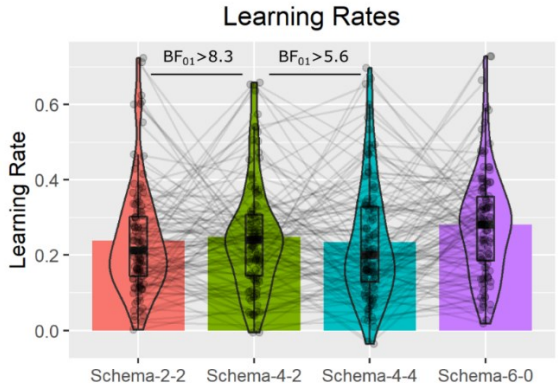


Supplementary Figure 8.6: The 2-parameter model estimates for learning rates for Experiment 1.
Heights of the bars indicate mean values.



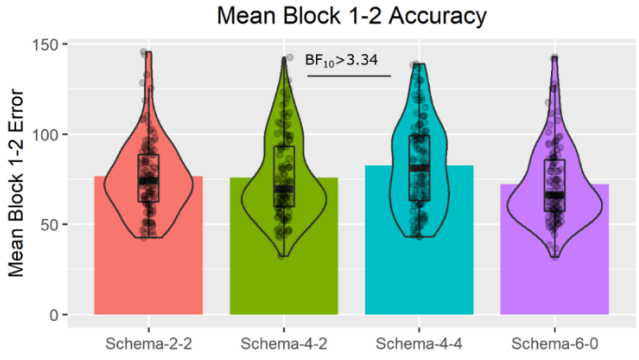
Supplementary Figure 8.7: The 2-parameter model estimates for Near vs Far-PA learning rates.
Heights of the bars represent mean values.

8.7 Experiment 2: learning rate estimates



Supplementary Figure 8.8: 2-parameter model learning rate estimates across conditions for Experiment 2.
Heights of the bars represent mean values.

8.8 Experiment 2: Blocks 1 and 2 Error combined



Supplementary Figure 8.9: Combined Block 1 and Block 2 error for the 4 conditions of Experiment 2.
Heights of the bars indicate mean values.