ACCOUNTING FOR ORDER AND MENTAL IMAGERY WITHIN MATHEMATICAL MODELS OF ASSOCIATION MEMORY

by

Jeremy J. Thomas

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Psychology
University of Alberta

Abstract

The general goal of this thesis was to uncover the computational characteristics of verbal association memory by focusing on two specific topics.

We first examined the role of mental imagery in association memory. One of the most effective ways to improve verbal association memory is to ask participants to form mental images of verbal stimuli. However, the functional role of mental imagery in cognition has been a subject of debate (Pearson & Kosslyn, 2015; Pylyshyn, 2002). An idea we test in the following work is if conscious mental imagery is an essential component of interactive imagery instructions. We tested this idea in chapter 2 by examining whether individual differences in both mental imagery vividness, or mental imagery ability predicted the benefit due to imagery instruction. We also examined how imagery instructions benefited a sub-population of individuals who report little to no imagery experience at all (aphantasia). We found that individual differences in visual imagery vividness and skill did not co-vary with the effectiveness of interactive imagery, and self-identified aphantasics benefited equally from imagery instructions. These results suggest that the visual image is not necessary for interactive imagery effects, and opens the possibility for alternative explanations of this effect, such as interactive imagery leading participants to encode more pair-unique representations of items.

Next, we examined memory for the constituent order of associations (AB versus BA). Existing mathematical models of association memory predict that associations are remembered with perfect order, or no order at all. However, empirical data indicates memory for the constituent-order of associations is moderate (Rehani & Caplan, 2011; Kato & Caplan, 2017). To help resolve this challenge to models, we first tested the possibility that imagery instructions could improve memory for constituent-order, perhaps to the levels predicted by perfect-order models. In chapter 2, we found that imagery instructions did not improve the ability to judge constituent-order (AB versus BA), nor the moderate relationship between order memory and association memory. This result increased the need to modify the mathematical models themselves. In chapter 3, we attempted to address this by extending convolution-based models, which normally store associations with no order, to store order in four ways. We evaluated these extended models against several behavioural benchmarks. We found that these extensions could account for moderate performance on order judgments; however, only one out of four could solve the additional benchmark of double function lists. This latter result suggests that to account for the full set of benchmark data, one needs to adopt specific assumptions about how constituent-order is represented in memory.

In the final chapter, we discussed how we might synthesize both of the major topics examined, considering our finding that interactive imagery instructions could not improve order recognition performance. This finding indicates that any account of imagery effects, or order memory, must also explain the in-variance of order memory to imagery instructions, providing an additional constraint for models. If a model can satisfy this constraint, it would simultaneously inform our computational account of both imagery effects and memory for constituent-order.

Preface

This thesis is an original work by Jeremy J. Thomas. All research projects contributing to this work received ethics approval from a University of Alberta Research Ethics Board, project name ”Structure of Human Memory” (Pro00105383). Chapter 2 was first published as: Thomas, J.J., Ayuno, K.C., Kluger, F.E., and Caplan, J.B. (in-press). The relationship between interactive-imagery instructions and association memory. *Memory & Cognition*, and reproduced with permission from Springer Nature. Jeremy J. Thomas conceived and designed this research with Jeremy B. Caplan. Jeremy J. Thomas conducted data analyses, interpreted the results, prepared the figures, wrote, and revised chapter 2. Kezziah C. Ayuno contributed to the conceptualization of, and data analysis for, experiment 2 in chapter 2. Felicitas Kluger contributed to writing in the introduction of chapter 2. Jeremy B. Caplan contributed to interpretation of results and edits to the manuscript. Jeremy J. Thomas conducted data analyses, interpreted the results, prepared the figures, wrote, and revised these chapters. Jeremy B. Caplan edited these chapters.

Acknowledgements

The following work would not have been possible without the network of colleagues, friends and family that has surrounded me in the past years. I would like to thank my supervisor, Dr. Jeremy Caplan. Beyond doing a superb job at teaching me the technical skills required to be a researcher, I have been greatly influenced by Dr. Caplan's excitement and love for science, which made working in his lab a continuous adventure. I would like to thank Dr. Alona Fyshe, and Dr. Peter Dixon for agreeing to be on my supervisory committee, along with all of the additional responsibilities that came with. I would like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC), and Department of Psychology of the University of Alberta for funding this research. There were many talented undergraduates that made our lab a vibrant and exciting place to work in. In particular, Kezziah Ayuno, who was a great help with some of the following work. I am also thankful for the friendships of past and present lab-members, Dr. Sucheta Chakravarty, Dr. Felicitas Kluger, and Tamari Shalamberidze. Thank you to my friends Joel and Kenny, my parents Jiju and Shirley, and my sister Jessica for all of their support. Finally, I would like to thank my wife Mikaela, who has been so encouraging at every step of the way.

Contents

List of Tables

List of Figures

Chapter 1

General introduction

A fundamental question about cognition is how the mind forms and remembers associations between pairs of stimuli (Kahana, 2012; Murdock, 1974). From simple face-name relationships, to remembering a sequence of ideas in a long speech, associations are an important component of many cognitive processes. One way psychologists have gained insight into human association memory is through verbal memory tasks with word pairs (study APPLE OVEN, given APPLE, recall OVEN). Data from word pair tasks have supported the development of a number of successful mathematical models, which have, in turn, given us more insight into the computational characteristics of human association memory (Kahana, 2012; Murdock, 1974).

In the following thesis, we continue with the general goal of uncovering the characteristics of association memory, but address two specific topics. First, one of the most effective ways to improve association memory is to ask participants to study verbal materials by forming mental images. However, as we elaborate in the present chapter, there is debate about whether the conscious experience of mental imagery corresponds to underlying cognitive representations (cf. imagery debate, Pearson & Kosslyn, 2015; Pylyshyn, 2002). This made us wonder whether the conscious experience of mental imagery is an essential component of the effectiveness of interactive imagery instructions, which is an idea we test in the following work.

Second, participants have a moderate ability to remember the constituent-order of associations (AB versus BA), but existing mathematical models of association memory either predict that associations are remembered with perfect order, or no order at all (Rehani & Caplan, 2011; Kato &

Caplan, 2017). This represents a significant mismatch between models and data that needs to be addressed. Can we solve this mismatch between existing models and empirical data, by directly modifying models themselves? Alternatively, given that imagery instructions are an effective way to improve association memory performance, might they also improve memory for the constituent order, perhaps even to the levels predicted by perfect-order models? We address both of these questions in the following work.

Before we begin addressing these specific topics, in this first chapter, we introduce the modelling framework that will be central to the following work. Then, we introduce some of the existing research that led us to ask our stated questions, including how the imagery debate (Pearson & Kosslyn, 2015; Pylyshyn, 2002) could speak to our interpretations of interactive imagery effect.

1.1 Representing items as sets of features

A common assumption held by many models of association memory is that any item, whether it be a word, picture or sound, are composed of discrete attributes or features that describe its unique characteristics (Kahana, 2012). For example, if we wanted to denote geometric shapes with different colours, we could imagine that colour would be one feature, while shape (square, circle, triangle etc.) would be another. We can implement this idea formally with vectors, where a red circle and a red triangle are denoted by $\mathbf{x} = \begin{bmatrix}1 & -1\end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix}1 & 1\end{bmatrix}$ respectively, where boldface denotes vectors. The first dimension in each vector denotes colour, where a value of 1 indicates red, and the second dimension denotes shape, where a value of 1 indicates triangle, and a value of -1 indicates circle. Vector representations are handy, because we use operations like the dot product to quantify the similarity between items, where larger dot products indicates more similarity. For example, by comparing the dot product $\mathbf{x} \cdot \mathbf{x} = 2$, and the dot product $\mathbf{x} \cdot \mathbf{y} = 0$, we can see that $\mathbf{x}$ is more similar to itself than to $\mathbf{y}$. Furthermore, if we assume vectors are normalized (have a length of 1), dot products are equivalent to the cosine of the angle between those vectors in Euclidean space, and vary from –1 to 1.

Although our example used vectors with two features, we would imagine that in the brain, items are represented with much more detail and require larger n-dimensional vectors. Drawing a closer analogy to the brain, we could imagine that n feature values in item vectors represent the firing rates of a population of n neurons, where certain neurons would fire at a higher rate, while others would fire less, unique to the item that is being represented. Also in our example, dimensions in each vector denoted specific properties of the item; however, in practice, most memory models keep features abstract and do not make assertions about what they represent, although there are some exceptions (Criss & Shiffrin, 2004a; Cox & Criss, 2017, 2020; Hintzman, 1988; Nairne, 1990).

1.2 Modelling associations between items

Now that we have a canonical representation of items, we can turn our attention to modelling the associations between them. Although there have been a number of ideas for this, in the following work we pay special attention to a class of models known as distributed memory models (Kahana, 2012).

One of the first models of this type was Anderson's (1970) matrix model, which encoded associations as matrix-outer products between pairs of items (Figure 1.1). This model is expressed as follows, $M = \mathbf{x}\mathbf{y}^\intercal$, where $\mathbf{x}$ and $\mathbf{y}$ denote n-dimensional item vectors, $\intercal$ denotes transpose, and M denotes an n×n memory matrix. M is a set of n×n connection weights encoding the association, or relating again to the brain, n×n synaptic strengths for the connections between two populations of neurons. Here we can see why matrix models are referred to as distributed models, because an association is encoded across all of the connection weights in the memory matrix. We can also obtain one matrix that encodes multiple associations by summing multiple memory matrices together, $M = \mathbf{x}\mathbf{y}^\intercal + \mathbf{a}\mathbf{b}^\intercal$, where $\mathbf{a}$ and $\mathbf{b}$ denote another pair of n-dimensional item vectors.

A key feature of matrix models is that they are content-addressable (Kahana, 2012). We can "cue" the memory matrix with any item vector stored in memory, which will recover an approximate of its associated item from the matrix, even when multiple other associations are stored in memory. This is implemented by multiplying a cue vector with the memory matrix. Assuming that all item vectors are normalized and approximately orthogonal to each other, $M\mathbf{y} \approx \mathbf{x} + \textit{noise}$. Based on a variety of parameters such as n, number of pairs stored in memory, and the similarity between stored item vectors, the recovered version of **x** is a somewhat noisy, imperfect approximation of the original item. While this leads the model to sometimes make errors, it also makes it robust in other interesting ways. For example, if the item vector used as a cue differs from the version stored in memory, the model can still retrieve a target item associated with an item that is most similar to the cue. However, if the messy cue vector is also similar to another item in memory, the model will become confused and simultaneously retrieve that association as well, which seems analogous to some of the similarity-based errors that human participants make (Anderson, 1970).

Another major family of distributed memory models is based on convolution. In these models, associations are encoded as the convolutions between item vectors, expressed as, $\mathbf{m} = \mathbf{x} * \mathbf{y}$, where $*$ denotes convolution and **m** denotes the memory vector (Figure 1.1). Similar to matrix models, multiple associations can be stored in the same trace by summing terms, $\mathbf{m} = \mathbf{a} * \mathbf{b} + \mathbf{x} * \mathbf{y}$. Convolution models are also content-addressable. We can cue the memory trace using circular correlation, which is the approximate inverse of convolution, $\mathbf{a} \# \mathbf{m} \approx \mathbf{b} + \textit{noise}$, where # denotes circular correlation. Like cued recall mechanisms in the matrix model, this operation is also quite noisy, and the model also makes similarity-based errors that are useful when modelling human memory. A number of models have applied the convolution-correlation framework to associative learning (Borsellino & Poggio, 1972; Gabor, 1969; Longuet-Higgins, 1968; Pribram, 1969; van Heerden, 1963), but TODAM (Murdock, 1982) and CHARM (Metcalfe Eich, 1982) were the first models to apply convolution to verbal association memory. Convolution has since been successful at modelling a wide range of psychological data, and some of its inherent properties such as its symmetry ($\mathbf{a} * \mathbf{b} \equiv \mathbf{b} * \mathbf{a}$), turn out to be a good match to certain characteristics of verbal memory (Kahana, 2002); however, as we elaborate below, also cause these models to generate erroneous predictions about other characteristics.

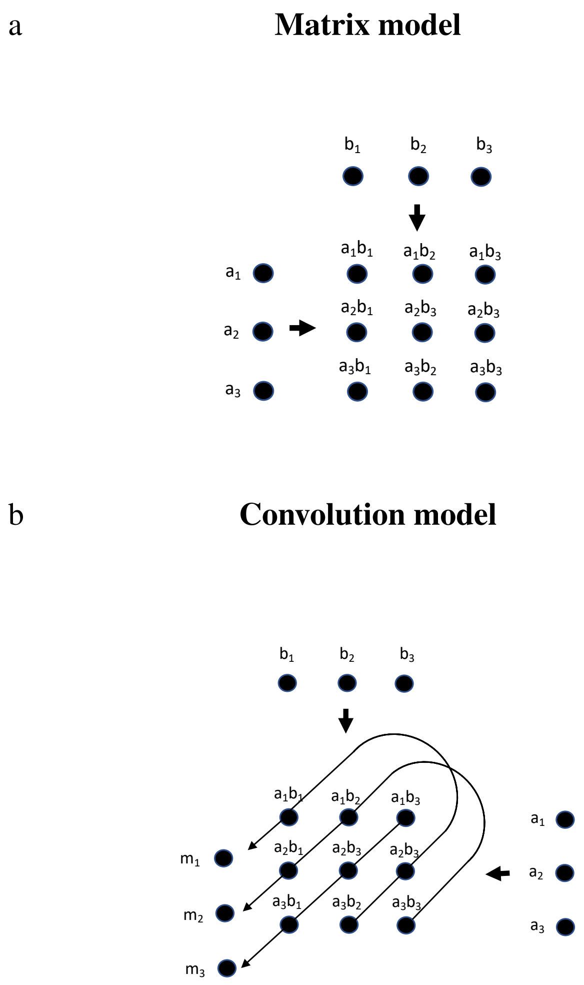

a
Matrix model
b_1 b_2 b_3
a_1b_1 a_1b_2 a_1b_3
a_1
a_2b_1 a_2b_3 a_2b_3
a_2
a_3b_1 a_3b_2 a_3b_3
a_3
b
Convolution model
b_1 b_2 b_3
a_1b_1 a_1b_2 a_1b_3
a_1
m_1
a_2b_1 a_2b_3 a_2b_3
a_2
m_2
a_3b_1 a_3b_2 a_3b_3
a_3
m_3

Figure 1.1: Illustrations of encoding operations for (a) matrix-based models: matrix-outer product between item vectors resulting a matrix, b) convolution-based models: the convolution between item vectors that results in a single vector, where $n = 3$, **a** and **b** are item vectors, and subscripts denote the index of the vector element. Vector **m** denotes the memory vector in convolution. Convolution-model illustrations are modified from figures in Plate (1995). As Plate (1995) noted, convolution is equivalent to a compression of the outer-product between two item vectors, which we depict here.

1.3 The problem of constituent-order

Although matrix models and convolution models are competing accounts of the same phenomena, and there have been arguments in favour of and against both (Pike, 1984; Murdock, 1985), both models have, overall, been successful at modelling a wide array of behavioural effects. However, a key difference between both is how they account for the constituent-order of associations (AB versus BA). This difference arises directly from the mathematical operations intrinsic to each model. Matrix outer-products are strictly non-commutative, meaning that $\mathbf{a}\mathbf{b}^\mathsf{T} \neq \mathbf{b}\mathbf{a}^\mathsf{T}$. By switching the presentation order of pair of items, the matrix model stores a completely different association in memory. On the other hand, convolution is a strictly commutative operation, $\mathbf{a} * \mathbf{b} \equiv \mathbf{b} * \mathbf{a}$, meaning that for a standard convolution model, presentation order has no effect on the association encoded in memory. Behavioural data contradicts both of these assumptions. Associations seem to be encoded with some moderate level of order that challenges all models (Kato & Caplan, 2017; Rehani & Caplan, 2011). As we elaborate in Chapter 3, existing attempts to modify matrix models to store associations with *less* order have been tried, but could not explain the fine details of the data. Instead, the approach we take in the following work is to modify convolution to store associations with more order (Chapter 3).

1.4 Mental imagery and association memory

There has been significant experimental work on uncovering the functional and theoretical importance of mental imagery in verbal association memory. Some of this work has examined explicit instructions to use mental imagery as a study strategy, where participants are instructed to form mental images with both words interacting together (e.g., for the pair DOG PIE, form a mental image of the DOG eating the pie). This strategy, known as interactive imagery, is one of the most effective ways to improve cued recall performance without training (Bower, 1970a; Bower & Winzenz, 1970; Dunlosky, Hertzog, & Powell-Moman, 2005; Paivio & Yuille, 1969; Paivio & Foth, 1970; Richardson, 1985, 1998). Participants instructed to use interactive imagery perform

$\sim 20 - 50\%$ better than participants instructed to use rote repetition (Bower & Winzenz, 1970; Bower, 1970b). The effectiveness of imagery instructions suggests that mental imagery can play a functional role in association memory performance. This idea is supported by related work focused on stimulus attributes such as imageability. Imageability is a subjective rating about how likely a word is to evoke mental images (Paivio, Yuille, & Madigan, 1968).[1] Multiple studies have found that word pairs with high imageability words are remembered better than word pairs with low imageability words (Paivio, Yuille, & Smythe, 1966; Paivio, Smythe, & Yuille, 1968; Paivio & Yuille, 1969). Additionally, participants report using imagery-related strategies more often when studying high-imageability words (Paivio, Smythe, & Yuille, 1968; Paivio & Yuille, 1969), suggesting that high-imageability words might be remembered better because they allow participants to use imagery-related memory strategies.

A classic explanation of both interactive imagery and imageability effects was Paivio's dual-coding theory (Paivio, 1969, 1971, 1986). Dual-coding theory proposes that there are two formats, verbal and imaginal, by which information can be stored in memory. Highly imageable words and/or verbal associations studied with mental imagery, are encoded with both a verbal and imaginal format, while low imageability words studied with non-imagery related strategies, are encoded with a verbal format. According to dual-coding theory, associations studied with interactive imagery are remembered more easily because both the verbal and imaginal format can be elicited at test.

However, the role of visual imagery in cognition has been the subject of a decades-long debate, which may be important to consider here. Kosslyn and colleagues argued that visual images correspond to a distinct *depictive* format by which information can be represented in the mind. In a depictive representation, the parts of that representation correspond to the parts of the physical object being represented, and the distances between the parts of the depictive representation correspond to the physical distances between the parts of the represented object (Pearson & Kosslyn,

[1] Imageability is highly correlated with stimulus *concreteness* (Paivio, Yuille, & Madigan, 1968), which is another attribute that we refer in this thesis. Although there are words with high imageability and low concreteness, and vice versa, many of the same effects are observed for both stimulus attributes, including, as we mention later, the high versus low advantage for cued recall performance.

2015). This view also assumes that these depictive representations co-exist with *descriptive*, or language-like representations in the mind (Pearson & Kosslyn, 2015). Pylyshyn's counter-position was that mental images do not necessarily provide evidence of a distinct depictive format, and that the experience of mental imagery may be epiphenomenal. Rather, we should first assume that all information, whether visual or verbal, is represented with a common format (Pylyshyn, 2002).

These contrasting ideas are relevant to our interpretation of interactive imagery effects. On one hand, interactive imagery instructions might be effective because they lead participants to encode associations with a distinctly (depictive) visual format alongside a descriptive verbal format. Alternatively, if imagery is epiphenomenal, as Pylyshyn (2002) argued, then one would expect that visual images are unnecessary for the underlying cognitive mechanisms leading to interactive imagery effects.

We used an individual differences approach to test these contrasting ideas. There are well-known individual differences both in self-reported mental imagery vividness (Marks, 1973; Zeman, Dewar, & Della Sala, 2015; Zeman et al., 2020), and skill in objectively-scored imagery-related tasks (Keogh & Pearson, 2018; Sanchez, 2019). Additionally, some individuals report aphantasia, or little to no experience of mental imagery at all (Zeman et al., 2015, 2020). We leveraged these individual differences to test contrasting hypotheses about the nature of interactive imagery instructions. If imagery instructions truly rely on visual imagery, this implies that individual ability to form vivid, or incredibly accurate visual images would allow participants to store additional detail in their stored image. In this case, participants with more visual imagery ability or vividness should benefit more from imagery instructions. Alternatively, if visual images do not correspond to underlying cognitive processes, then we should expect that additional detail or accuracy in the visual image should not have any effect. We elaborate on how we tested these competing hypotheses in Chapter 2.

1.5 The effect of interactive imagery on memory for constituent-order

Bringing these threads together, although mathematical models have typically not incorporated the effects of visual imagery, imagery instructions may cause qualitative differences in memory that are relevant for mathematical models.

More specifically, image-like representations may allow participants to store additional detail about the constituent-order of the association (AB versus BA). Pearson and Kosslyn (2015) suggested that depictive image-like representations may allow us to recover information at test that we did not directly focus on at study. For example, if you encode a visual image of a kitchen filled with appliances, perhaps at test you could retrieve the image of the kitchen and use it to recall the spatial organization of the various appliances in the room. Using a similar strategy, participants may be able to use stored images to recover the constituent-order of an association with great accuracy, which would be relevant for existing challenges to models. If imagery instructions lead participants to store associations with images, and participants can use these images to store the constituent-order of a pairing with great detail, this might improve order memory even to the levels predicted by perfect-order, matrix models. If this were the case, then matrix models would actually be supported under some conditions. This is an idea that we also test in the following work (Chapter 2).

1.6 Summary of research goals

Based on our review of the literature and analysis of existing mathematical models, we identified the following research questions:

1. Is the conscious experience of mental imagery and/or mental imagery skill necessary to benefit from interactive imagery instructions?

2. Can interactive imagery instructions lead participants to store associations with more order?

3. Is there a way to modify existing models to store moderate levels of order, and better account for empirical data?

To address question 1, in chapter 2 we present three behavioural experiments. Each of these experiments tested association memory with cued recall (study AB, given A, recall B), with a pre/post imagery instruction manipulation, where we first measured baseline memory performance for each participant, then administered imagery instructions halfway through the lists. This allowed us to measure performance uninstructed memory performance and performance after imagery instruction for each participant, giving us closer sense of the effect of imagery instructions. At the end of the session in each experiment we measured individual differences in imagery vividness with the Vividness of Visual Imagery Questionnaire (Marks, 1973). In experiment 1 and 3 we also measured objective imagery skill with the Paper Folding Task (French, Ekstrom, & Price, 1963). Both of these tasks are described in more detail in the following chapter. We computed the correlation between these two individual difference measures and memory performance to test address question 1.

To address question 2, experiments presented in chapter 2 included an order recognition task immediately after cued recall tests, where participants judged the constituent-order of pairs (experiments 1 and 2).

To address question 3, in chapter 3 we proposed four extensions of symmetric, order-absent models of association memory that could address question 3. This section begins with an overview of the challenging empirical benchmarks regarding memory for constituent-order that each model must overcome. We then provide formal expressions for the specific extensions of convolution that we are proposing. Then, using a combination of Monte Carlo simulations, and algebraic arguments, we evaluate each model extension against these empirical benchmarks.

Chapter 4 will present a summary of the previous chapters, and synthesis of the main themes.

Chapter 2

The relationship between interactive imagery instructions and association memory

Abstract

Interactive imagery, one of the most effective strategies for remembering pairs of words, involves asking participants to form mental images during study. We tested the hypothesis that the visual image is, in fact, responsible for its memory benefit. Neither subjectively reported vividness (all experiments) nor objective imagery skill (experiments 1 and 3) could explain the benefit of interactive imagery for cued recall. Aphantasic participants, who self-identified little to no mental imagery, benefited from interactive imagery instructions as much as controls (experiment 3). Imagery instructions did not improve memory for the constituent-order of associations (AB versus BA), even when participants were told how to incorporate order within their images (experiments 1 and 2). Taken together, our results suggest that the visual format of images may not be responsible for the effectiveness of the interactive imagery instruction and moreover, interactive imagery may not result in qualitatively different associative memories.

2.1 Introduction

One of the best known ways to increase memory for word pairs (e.g., study APPLE-OVEN, when presented APPLE, recall OVEN), is to instruct participants to form a mental image of the two words interacting (Bower, 1970a; Bower & Winzenz, 1970; Dunlosky et al., 2005; Paivio & Yuille, 1969; Paivio & Foth, 1970; Richardson, 1985, 1998). For example, "imagine an APPLE cooked inside an OVEN, in your mind's eye." Participants who receive interactive imagery instructions perform significantly better at cued recall than participants given no strategy instruction (Richardson, 1985, 1998), and $\sim 20-50\%$ higher cued recall accuracy than participants instructed to use rote repetition (Bower & Winzenz, 1970; Bower, 1970a). Bower and Winzenz (1970) and Paivio and Foth (1970) found that interactive imagery instructions could even outperform comparable verbally mediated instructions (e.g., form a sentence with both words) for concrete word pairs, although Dunlosky et al. (2005) found these instructions were comparable. At face-value, interactive imagery instructions might cause participants to literally construct rich visual representations, directly improving memory in this way (Yates, 1966). However, this hypothesis is hard to test because visual imagery cannot be directly observed. Here we examine the effect of interactive imagery instructions with two main approaches. First, we test the visually relevant characteristics of the imagery instruction and individual differences characteristics of the participants. Second, we ask whether interactive imagery changes the formal nature of the representation; specifically, whether or not constituent-order (knowledge that it was APPLE–OVEN, not OVEN–APPLE) is coupled with memory for the pairing, itself.

Testing for visual-imagery characteristics of associations formed through interactive imagery

One way to interrogate how visual imagery functions is to exploit individual differences. There is large individual variability in the subjective experience of mental imagery (Marks, 1973; Zeman et al., 2015, 2020) and objectively scored imagery/visuospatial tasks (Keogh & Pearson, 2018; Sanchez, 2019; Zeman et al., 2010). If the visual image, itself, is fundamental to the benefit of interactive imagery, one would expect that imagery instructions may benefit individuals with vivid or accurate mental imagery more than those with poor mental imagery. Alternatively, visual imagery may be epiphenomenal (Pylyshyn, 2002), implying that individual differences in mental imagery should not relate to objective memory performance. Our three experiments test the hypothesis that both mental imagery vividness and skill determine how much an individual benefits from interactive imagery instructions.

There is considerable support for a central role of imagery in association-memory. Instructions to use interactive imagery produces higher cued recall than without imagery instructions, and associations involving words higher in imageability are remembered better (Bower, 1970a; Bower & Winzenz, 1970; Paivio, Smythe, & Yuille, 1968; Paivio & Yuille, 1969; Paivio, 1969; Paivio & Foth, 1970). Beyond memory for word pairs, ancient texts claim that forming vivid images can improve memory of various kinds (Foer, 2011; Gesualdo, 1592; Yates, 1966). For example, when using the Method of Loci, a popular technique for ordered lists, skilled memorizers report forming mental images of to-be-remembered items in various locations (e.g., Maguire, Valentine, Wilding, & Kapur, 2003).

Common advice by skilled memorizers is that vivid imagery is important for the efficacy of mnemonic strategies (e.g., Foer, 2011; Konrad, 2013; Müller et al., 2018). To test this idea, Sanchez (2019) measured individual differences in imagery/visuospatial skill with the Cube Comparisons Task (CCT; a mental rotation task), and the Paper Folding Task (PFT; judging the outcome of multiple folds and hole-punches of a paper) (French et al., 1963), and examined the correlation to memory performance. In Sanchez' (2019) study, aggregate CCT and PFT performance correlated with serial recall performance for participants who were instructed to use the Method of Loci, but not for participants who were given a control instruction. However, three studies did not find a significant relationship between Vividness of Visual Imagery Questionnaire (VVIQ; Marks, 1973) and successful use of the Method of Loci (Kliegl, Smith, & Baltes, 1990; Kluger, Oladimeji, Tan, Brown, & Caplan, 2022; McKellar, Marks and Barron, cited as in-preparation by Marks, 1972).

In light of these variable findings, we included the VVIQ (all experiments) and PFT (experiments 1 and 3) to assess subjective quality of imagery and objective imagery ability, respectively. The hypothesis that the construction of a visual image is central to the success of interactive imagery instructions implies that either or both the VVIQ and PFT should covary with cued recall accuracy. Alternatively, interactive imagery effects may not depend on vivid or accurate mental images, or perhaps, do not require any conscious experience of mental imagery at all.

To further test the hypothesis that visual imagery is vital for the benefits of interactive imagery, we tested people with the phenomenon of aphantasia, extremely low or non-existent self-reported ability to form voluntary mental images. Current interest in aphantasia originated with patient MX (Zeman et al., 2010), who, after undergoing coronary angioplasty, reported a complete inability to form mental images. MX exhibited completely intact performance in imagery/visuospatial related tasks. However, closer examination of behaviour and brain activity suggested MX was applying distinct verbal/symbolic strategies to complete tasks typically thought to require mental imagery. Other studies have examined larger populations of self-reported aphantasics who rate significantly low vividness (Zeman et al., 2015), report worse autobiographical memory and difficulty with recognizing faces (Zeman et al., 2020). Specific to memory, Bainbridge, Pounder, Eardley, and Baker (2021) examined the ability of aphantasics to draw photographs of rooms in a house from memory. Aphantasics were not different from controls in copying a presented image, indicating no deficits to their perceptual ability. Interestingly, aphantasics remembered fewer objects than controls, but for the objects they could remember, they reproduced their spatial arrangement at the same level as controls. These results indicated that aphantasics had specific deficits to object, but not spatial memory. If the visual image is the necessary mechanism by which interactive imagery instructions increase cued recall accuracy, aphantasics should show no such advantage (experiment 3).

Interactive imagery and the formal properties of associations We could find no formal implementation of imagery in any mathematical model of association-memory. However, image-based associations could differ in their qualitative or formal characteristics, which might be meaningful from a mathematical modelling perspective. One hypothesis about the relationship between imagery and the formal characteristics of association-memory emerged while reviewing existing models, as we now elaborate.

Mathematical models make starkly different predictions about memory for the constituent-order of associations (AB versus BA) (Kato & Caplan, 2017), a memory task that has only begun to be investigated experimentally. Matrix-based models (Anderson, 1970) and concatenation-based models (Hintzman, 1984; Shiffrin & Steyvers, 1997), which we now refer to as perfect-order models, encode associations with non-commutative operations, and consequently predict that order is remembered perfectly given that the association itself is intact. Convolution-based models (Kelly, Blostein, & Mewhort, 2013; Murdock, 1982; Metcalfe Eich, 1982; Plate, 1995), in contrast, are based on commutative operations that completely discard order (and see Cox & Criss, 2017, 2020 and Criss and Shiffrin's 2005 model, which also disregard order). In these models, which we now refer to as order-absent models, information for order, if present, must be provided by some other term, predicting that the ability to remember the constituent-order will be unrelated to remembering the pairing itself. Kato and Caplan (2017) found no evidence for either of these predictions. In their study, word pairs were tested with cued recall, and then, an order recognition task, where participants had to recognize whether a probe was in the correct order (AB), or reversed (BA) (Greene & Tussing, 2001; Kounios, Bachman, Casasanto, Grossman, & Smith, 2003; Kounios, Smith, Yang, Bachman, & D'Esposito, 2001; J. Yang et al., 2013). Challenging both perfect-order and order-absent models, they found a significant correlation between order recognition and cued recall performance; however, this correlation was significantly smaller than a control correlation (with associative recognition), suggesting associations are not stored with perfect order, nor are they completely order-absent. If we take imagery at face-value, it seems plausible that a visual image could provide an effective means of incorporating order, such as left-to-right within the image, or top-to-bottom. This might be just the thing that participants are missing in their spontaneously adopted strategies. So in addition to increasing memory accuracy, interactive imagery instructions might help participants incorporate order, and render the association non-commutative like in a perfect-order model. This was our first hypothesis. The alternative hypothesis is that imagery is simply a good "hook", engaging participants better in the task, but otherwise invoking the same associative mechanism as in conditions without imagery instructions. This hypothesis leads to the prediction that the relationship between order and the association itself will be unchanged with interactive imagery instructions. We tested these two hypotheses in experiments 1 and 2 with order recognition subsequent to cued recall for all studied pairs in one group, and as a control, associative recognition in another group.

Summary of experiments In all experiments, participants studied lists of eight word-pairs followed by cued recall. First we obtained a baseline measure of memory with no strategy instructions, then participants were given imagery instructions (all experiments), or a filler instruction (experiment 1). To test the hypothesis that visual images are necessary for memory benefit due to interactive imagery, and that individual differences in imagery ability/vividness should predict memory benefit, vividness was assessed with the VVIQ in all experiments, and imagery skill was assessed with the PFT in experiments 1 and 2. Experiment 3 applied a stronger test of the visual imagery hypothesis by recruiting aphantasics. In experiments 1 and 2, we also tested the hypothesis that imagery could provide a way for participants to incorporate order and generate associations that are more non-commutative (like a matrix model). Cued recall was followed by either order or associative recognition, to test the relationship between constituent-order and memory for the pair, itself. The prediction is that imagery instructions will increase order recognition and moreover, its relationship to cued recall. Finally, we also include supplementary materials with additional analyses.

2.2 Experiment 1

2.2.1 Methods

Participants

Participants enrolled in introductory psychology courses at the University of Alberta ($N = 227$) participated for partial course credit. Participants were required to have learned English before the age of six, have normal or corrected-to-normal vision, and be comfortable typing. Participants chose one of 15 testing rooms in order of arrival, blind to condition. One participant was excluded from analyses for not completing the experiment within the allotted 50 minutes. Procedures in all experiments were approved by a University of Alberta ethical review board.

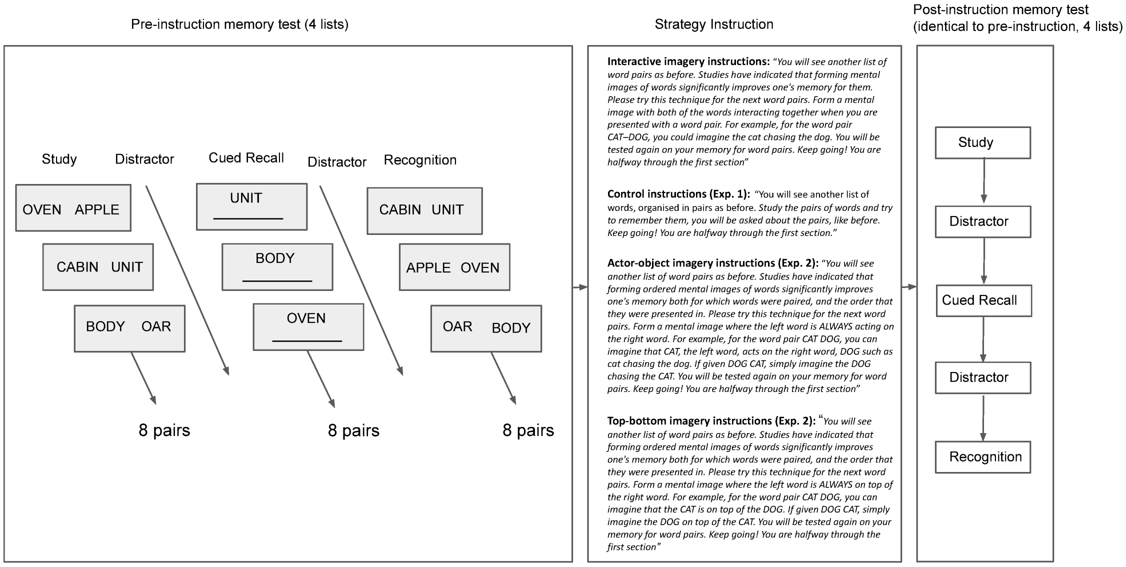
Groups There were two main experimental groups. The imagery group ($N = 113$) received interactive imagery instructions halfway through the word lists, and the control group ($N = 114$) received filler instructions halfway through the lists (Figure 2.1). Each experimental group was further subdivided into two conditions. Following cued recall, one condition performed order recognition ($N = 57$ and 56 for imagery and control, respectively), and the other condition performed associative recognition ($N = 56$ and 58, respectively). For analyses involving only cued recall, these conditions were collapsed within the imagery group and control group. For all analyses involving recognition tasks, these conditions were separated and named, accordingly, control-order recognition, control-associative recognition, imagery-order recognition, and imagery-associative recognition.

Materials

Stimuli were the 478 nouns from the Toronto Word Pool (Friendly, Franklin, Hoffman, & Rubin, 1982), four to eight letters and spanning the full ranges of concreteness mean (SD) = 5.32 (1.32), and with frequency = 62.47 (82.45) per million (Kucera & Francis, 1967). Words were assigned to pairs and lists with the computer's random number generator. Study pairs, cued recall and recognition test probes were presented in uppercase, white, Courier bold font.

Procedure

The experiment was run in Python, in conjunction with the Python Experiment-Programming Library (Geller, Schleifer, Sederberg, Jacobs, & Kahana, 2007), for the first cohort of participants. Because software updates made lab computers incompatible with PyEPL, we ran the second cohort in a MATLAB port, written with the PsychToolBox experiment programming extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997), and the CogToolBox Library (Fraundorf et al., 2014). Illustrated in Figure 2.1, the session included study of word pairs, cued recall, followed by order or associative recognition tests, repeated for eight study sets, with five



Interactive imagery instructions: *"You will see another list of word pairs as before. Studies have indicated that forming mental images of words significantly improves one's memory for them. Please try this technique for the next word pairs. Form a mental image with both of the words interacting together when you are presented with a word pair. For example, for the word pair CAT–DOG, you could imagine the cat chasing the dog. You will be tested again on your memory for word pairs. Keep going! You are halfway through the first section"*

Control instructions (Exp. 1): "You will see another list of words, organised in pairs as before. *Study the pairs of words and try to remember them, you will be asked about the pairs, like before. Keep going! You are halfway through the first section."*

Actor-object imagery instructions (Exp. 2): *"You will see another list of word pairs as before. Studies have indicated that forming ordered mental images of words significantly improves one's memory both for which words were paired, and the order that they were presented in. Please try this technique for the next word pairs. Form a mental image where the left word is ALWAYS acting on the right word. For example, for the word pair CAT DOG, you can imagine that CAT, the left word, acts on the right word, DOG such as cat chasing the dog. If given DOG CAT, simply imagine the DOG chasing the CAT. You will be tested again on your memory for word pairs. Keep going! You are halfway through the first section"*

Top-bottom imagery instructions (Exp. 2): "*You will see another list of word pairs as before. Studies have indicated that forming ordered mental images of words significantly improves one's memory both for which words were paired, and the order that they were presented in. Please try this technique for the next word pairs. Form a mental image where the left word is ALWAYS on top of the right word. For example, for the word pair CAT DOG, you can imagine that the CAT is on top of the DOG. If given DOG CAT, simply imagine the DOG on top of the CAT. You will be tested again on your memory for word pairs. Keep going! You are halfway through the first section"*

Figure 2.1: There were a total of eight lists in experiments 1 and 2. Halfway-through the lists participants either received imagery or control instructions in experiment 1, and either imagery, actor-object or top-bottom instructions in experiment 2. All participants in experiment 3 received imagery instructions. Experiment 3 had a similar design, but without associative or order recognition trials after cued recall, a total of ten lists, and all participants received imagery instructions.

trials of a mathematical distractor task between study, cued recall and recognition sets. Given that Kato and Caplan (2017) found that initial cued recall tests affected subsequent recognition tests but did not change the coupling of order with association-memory, we tested every pair initially with cued recall (as in experiment 1 of Kato & Caplan, 2017) to maximize the data yield (and see page 47). Interactive imagery instructions or control filler instructions were administered after the fourth list in a pretest (Lists 1–4)/posttest (Lists 5–8) design, allowing us to check for equal baseline performance (pre-instruction), and get a closer estimate of the true effect of imagery instructions above baseline. Participants then completed the VVIQ and the PFT. Halfway through data collection, a section was added after the PFT, where participants were asked to rate how often they used interactive imagery, and then asked to type a free-form response about their strategy use, reported on page 48

Practice list Participants performed one practice list excluded from analyses, at the beginning of the session, during which they were walked through the tasks.

Study phase For each list, participants viewed eight pairs in sequence. The two words in a pair were presented side by side, centered on the screen, for 2850 ms, with a 150-ms inter-pair blank.

Distractor Interleaved between study, recall and recognition, participants were administered a math distractor task. Participants had to solve the sum of three digits, randomly drawn from two to eight within 5000 ms followed by a 200-ms blank inter-trial interval. Participants typed their response, which was displayed on the screen, and upon pressing ENTER, the colour of the response digit changed to gray, to show the response registered, and the 200-ms inter-trial interval was initiated after the 5000-ms response interval elapsed.

Cued recall Each studied pair was tested once with cued recall. Direction of cued recall (forward, APPLE–?, or backward, ?–OVEN) was counterbalanced (Python version: across all lists except the practice; MATLAB version: within each list). The cue word was presented in centrally with a centered response line underneath, regardless of the direction of cued recall. The letters appeared on the line as the participant typed, submitting the word with the ENTER key. The next cued recall trial started 750 ms later. ENTER was only accepted once more than two letters were typed, to reduce participants speeding through. In the Python version, if participants did not press ENTER within 15,000 ms, the trial ended, was scored incorrect, and the next cued recall trial was presented. In the MATLAB version, this time-limit was removed.

Recognition Two probe words were presented side by side centrally, as in the study phase. In order recognition, participants judged if a presented probe was intact (e.g., OVEN APPLE) or reverse (e.g., APPLE OVEN). In associative recognition, participants judged whether a presented probe was intact (e.g., OVEN APPLE) or recombined (e.g., OVEN BUTTON). Key 1 was assigned to intact and key 2 was assigned to reverse or recombined. Other keys were ignored. Recombined probes were only rearranged with other pairs within the current list, and a pair probed with an intact probe was never used to create a recombined probe. Pairs were tested in pseudo-random order. In the Python version, the number of intact and lure (reverse or recombined) probes were counterbalanced over all analyzed lists (excluding practice). In the MATLAB version, trials were counterbalanced over all lists including the practice list.[1] In the Python version, the trial was

[1] Due to programming error, counterbalancing was slightly unbalanced for associative recognition in the MATLAB version. When one recombined trial was randomly assigned to given list, it did not have another recombined pair to

aborted after 15,000 ms. Rather than score these timed-out trials as incorrect, they were omitted from analyses (two trials in all, both in control-associative participants). To prevent missing data, the 15,000 ms timeout limit was removed in the MATLAB version. The next recognition trial started after a 750-ms blank screen.

Vividness of Visual Imagery Questionnaire Participants completed a computerized version of the Visual Vividness of Imagery Questionnaire (Marks, 1973), which asks participants to imagine four scenes. A description of each scene was displayed on the screen, followed by instructions to imagine four items within the scene and to rate vividness on a scale from one (perfectly vivid imagery), to five (no image at all) using the number keys. To indicate the response registered, the choice changed to green for 1000 ms, immediately followed by the next item. VVIQ score was the sum of these ratings, ranging from 16 (perfectly vivid imagery) to 80 (no image formed at all).

Paper Folding Task Participants completed a computerized version of the PFT (French et al., 1963), consisting of 20 questions increasing in difficulty. Each question was a series of images that depicted a piece of paper being folded successively and then hole-punched. The question was displayed to the left of a central vertical line, and five possible choices were displayed to the right, selected with the keys 1–5. The chosen option was highlighted in green for 1000 ms, immediately followed by the next question. Mean accuracy and response time were analyzed.

Distribution of VVIQ ratings and PFT ratings Distributions of VVIQ ratings and PFT scores aligned with previous studies (Table 2.1).

Analyses

To check null effects, we include Bayesian analyses (with uniform priors) run in JASP (JASP Team, 2021). The Bayes Factor is a ratio of evidence, where by convention, when $BF_{10} > 3$, the effect is considered supported, and when $BF_{10} < 0.3$, the effect is considered more consistent with the null. For ANOVAs, $BF_{\text{inclusion}}$, which summarizes across all factorial models and quantifies whether each model fits better with the main effect or interaction included versus excluded. We measured order and associative recognition with $d' = z(\text{hit rate}) - z(\text{false alarm rate})$. Whenever

exchange words with, and appeared as an intact trial. The occurrence of this error was rare, with 11 participants having one extra intact trial, and one participant having two extra intact trials.

Table 2.1: M(SD) (Means and standard deviations) of VVIQ ratings for each group in experiments 1, 2 and 3, and PFT scores in experiment 1, and 3, along with population estimates for VVIQ ratings from McKelvie's (1995), and PFT scores in the control and method of loci group in Sanchez (2019).

Experiment and Group	VVIQ Rating	PFT score
Sanchez (2019) method of loci group	N/A	12.52 (2.59)
Sanchez (2019) control group	N/A	11.87 (3.30)
McKelvie (1995) VVIQ population estimate	36.9 (11.07)	N/A
Experiment 1: Imagery-order recognition sub-condition	31.8 (10.86)	13.02 (4.06)
Experiment 1: Imagery-associative recognition sub-condition	32.9 (8.94)	13.70 (3.87)
Experiment 1: Control-order recognition sub-condition	32.5 (8.39)	13.14 (3.75)
Experiment 1: Control-associative recognition sub-condition	32.7 (9.73)	13.83 (4.54)
Experiment 2: Actor-object-order recognition sub-condition	36.2 (12.62)	N/A
Experiment 2: Actor-object-associative recognition sub-condition	36.2 (10.07)	N/A
Experiment 2: Standard-imagery-order recognition sub-condition	36.3 (10.52)	N/A
Experiment 2: Standard-imagery-associative recognition sub-condition	35.2 (8.07)	N/A
Experiment 2: Top-bottom-order recognition sub-condition	36.9 (11.92)	N/A
Experiment 2: Top-bottom-associative recognition sub-condition	35.7 (11.35)	N/A
Experiment 3: Consistent aphantasic group	61.0 (18.06)	12.40 (4.61)
Experiment 3: Consistent non-aphantasic group	38.1 (13.89)	13.15 (4.37)
Experiment 3: Inconsistent responder group	44.7 (15.67)	11.98 (4.50)

hit or false alarm rate were zero or one, one-half an observation was added or subtracted to avoid infinities.

2.2.2 Results and discussion

Cued recall We replicated the interactive-imagery advantage for cued recall. A mixed ANOVA on cued recall accuracy (Figure 2.2), with design Group (imagery, control group) × Instruction phase (pre-instruction, post-instruction), returned significant main effects of Instruction phase, $F(1,225) = 110.79$, $MSE = 2.91$, $p < .001$, $\eta_p^2 = 0.33$, $BF_{\text{inclusion}} > 1000$, and Group, $F(1,225) = 4.92$, $MSE = 0.41$, $p = .03$, $\eta_p^2 = 0.02$, $BF_{\text{inclusion}} > 1000$; however, the interaction was also significant, $F(1,225) = 41.5$, $MSE = 1.09$, $p < .001$, $\eta_p^2 = 0.16$, $BF_{\text{inclusion}} > 1000$. Simple effects found no difference between groups pre-instruction ($p = .19$, $BF_{10} = 0.33$), but significantly higher accuracy for the imagery group post-instruction ($p < .001$, $BF_{10} > 1000$). Additionally, for both groups, accuracy significantly increased post-instruction (both $p < .001$, $BF_{10} > 33$). Thus, perhaps due to practice effects, the control group moderately improved as the experiment progressed; however, the imagery group performed significantly better in the post-instruction phase, and exhibited a greater improvement from baseline compared to the control group.[2]

Associative and order recognition A mixed ANOVA on associative recognition d' (Figure 2.3), with design Group (imagery-associative recognition, control-associative recognition) × Instruction phase (pre-instruction, post-instruction) returned a non-significant main effect of Group ($p = .25$, $BF_{\text{inclusion}} = 612.89$)[3], a significant main effect of Instruction phase, $F(1,112) = 38.13$, $MSE = 22.79$, $p < .001$, $\eta_p^2 = 0.25$, $BF_{\text{inclusion}} > 1000$, and a significant interaction Group × Instruction phase, $F(1,112) = 21.24$, $MSE = 13.29$, $p < .001$, $\eta_p^2 = 0.17$, $BF_{\text{inclusion}} > 1000$. Simple effects revealed a non-significant group difference in performance pre-instruction ($p = .14$, $BF_{10} = 0.54$), but the imagery-associative recognition condition performed significantly better post-instruction ($p < .001$, $BF_{10} = 31.12$). Additionally, the imagery-associative recognition condition improved post-instruction ($p < .001$, $BF_{10} > 1000$), but the control-associative recognition condition did not significantly improve ($p = .16$, $BF_{10} = 0.37$). These analyses indicate that imagery instructions

[2] Expanding on these findings, we also found evidence that imagery instructions were most beneficial for participants with poor baseline performance (page 50).

[3] A non-significant effect can have strong evidence in a Bayesian analysis because JASP's implementation of Bayesian model selection refuses to consider models including interactions without the main terms. Thus, if there is strong evidence for the interaction, it will also return strong evidence for the main terms included in interactions .

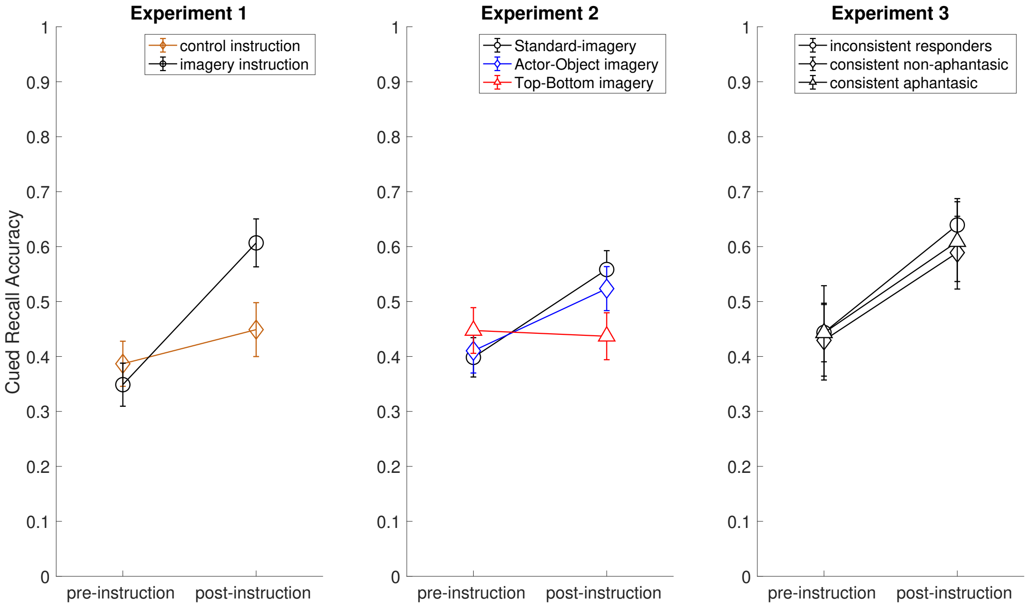


Figure 2.2: Pre- and post-instruction cued recall accuracy for all three experiments. (Left) In experiment 1, the imagery group received instructions to use interactive imagery halfway through the word lists. The control group was simply instructed to continue with the experiment. (Middle) In experiment 2, participants either received standard-imagery, actor-object imagery, or top-bottom imagery instructions. (Right) In experiment 3, all participants received imagery instructions. Error bars represent 95% confidence intervals based on standard error of the mean.

substantially improved associative recognition performance over control instructions.

An ANOVA with the same design, on order recognition d' (Figure 2.3) returned non-significant, favoured null main effects of both factors (both $p > .2$, $BF_{\text{inclusion}} < 0.3$). The interaction Group × Instruction phase nearly reached significance, $F(1,111) = 3.90, MSE = 1.61, p = .051, \eta_p^2 = 0.03$, although the Bayesian analysis favoured the null ($BF_{\text{inclusion}} = 0.26$). Nonetheless, we cautiously followed up on the interaction with simple effects. The control-order recognition group performed significantly worse post-instruction ($p = .01$, $BF_{10} = 3.07$), while the imagery-order recognition group did not exhibit any significant change ($p = .65$, $BF_{10} = 0.16$). Additionally, the group difference in performance was not significant pre-instruction ($p = .06$, $BF_{10} = 0.98$), or post-instruction ($p = .80$, $BF_{10} = 0.21$). In sum, imagery instructions did not improve order recognition performance, but may have acted against a performance decrease observed in the control-order recognition group.

The relationship among mental imagery skill, vividness, and the effectiveness of interactive imagery instructions Next, we asked if any individual difference measure would explain individual differences in memory performance (Tables 2.5–2.7). Correlations between VVIQ ratings and cued recall accuracy were all non-significant and either were, or were nearly, supported null effects (all $p > .09$, $BF_{10} < 0.45$), and likewise for order recognition (all $p > .15$, $BF_{10} < 0.46$). VVIQ ratings significantly correlated with post-instruction associative recognition performance in the imagery-associative recognition condition, $r(54) = -.44$, $p < .001$, $BF_{10} = 44.10$, but this correlation was not significant post-instruction for control-associative recognition group, $r(56) = -.04$, $p = .78$, $BF_{10} = 0.17$; and these correlations differed significantly (Fisher test, $p = .024$). Thus, individual differences in mental imagery vividness explained differences in associative recognition performance under interactive imagery conditions,[4] but could not explain the interactive imagery advantage for cued recall.

PFT accuracy exhibited significant, positive correlations with nearly all memory tasks, and not only with memory performance in the imagery group (Tables 2.5–2.7). Although the tables show some exceptions, our results, particularly the presence of pre-instruction correlations, suggest that PFT accuracy does not specifically relate to interactive imagery, and may have either reflected a general factor such as motivation, task engagement or a distinct cognitive process such as working

[4] The significant correlation between VVIQ ratings and post-imagery instruction associative recognition d' was not replicated in experiment 2, thus, we do not consider this a robust finding and do not discuss it in the general discussion.

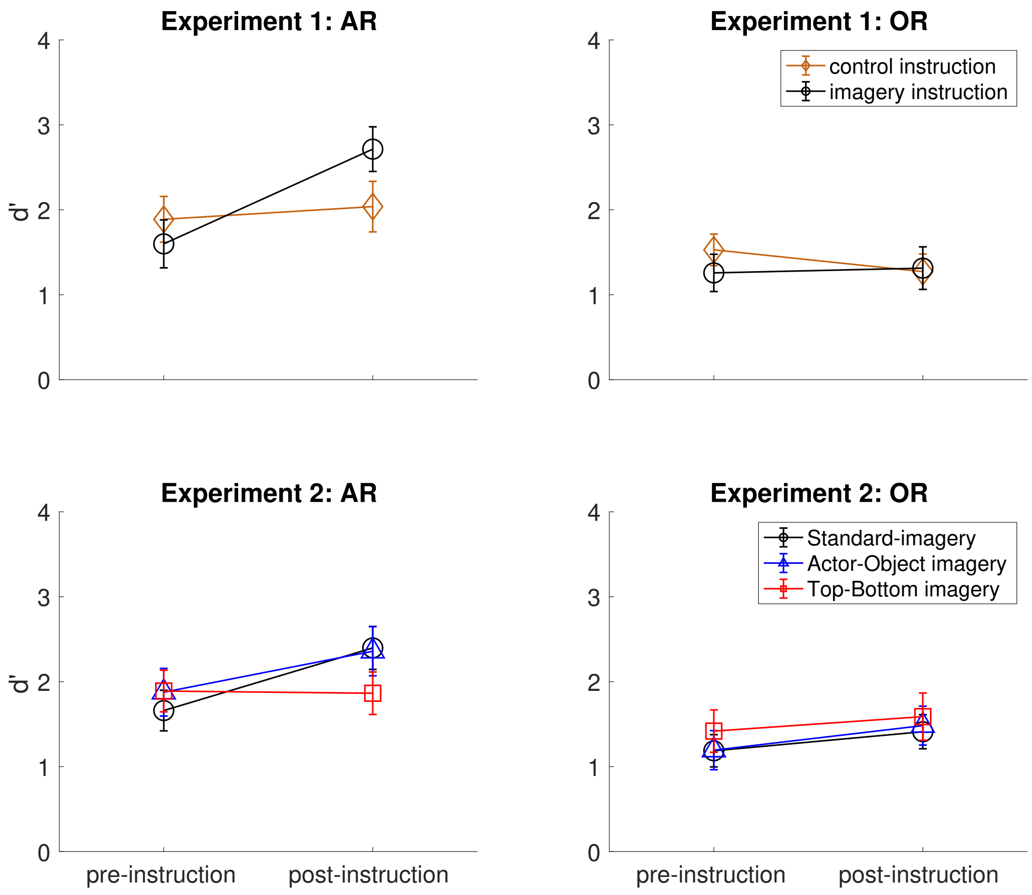


Figure 2.3: Pre- and post-instruction order (OR), and associative recognition (AR) performance for experiment 1 and 2. In experiment 1, participants either received standard imagery instructions or control instructions. In experiment 2, participants received either standard-imagery, top-bottom imagery, and actor-object imagery instructions. Error bars represent 95% confidence intervals based on standard error of the mean.

memory.

PFT response time was not significantly related to the memory measures apart from a significant positive correlation with post-instruction cued recall accuracy, $r(111) = .27$, $p = .004$, $BF_{10} = 7.49$, and post-instruction associative recognition performance, $r(54) = .32$, $p = .017$, $BF_{10} = 2.74$, both in the imagery group. If *longer* PFT response times indicate worse performance, these correlations would be counter-intuitive. A simpler interpretation is that longer PFT latencies are a consequence of greater general effort or engagement (a successful speed–accuracy trade-off) rather than mental imagery skill. Thus, the pattern argues against the idea that mental imagery accuracy or skill is required for the memory benefit.[5]

The relationship of order recognition to cued recall Figure 2.15 plots log-odds transformed cued recall accuracy versus both order recognition and associative recognition d', for both imagery and control groups. Pre-instruction, the associative recognition–cued recall correlations (imagery: $r(56) = .86$, $p < .001$, control: $r(56) = .83$, $p < .001$), were larger than the order recognition–cued-recall correlations (imagery: $r(55) = .43$, $p < .001$, control: $r(54) = .46$, $p < .001$). The difference in correlations was significant for both groups pre-instruction (Fisher tests, imagery: $p < .001$, control: $p < .001$). This pattern persisted post-instruction; associative recognition-cued recall correlations (imagery: $r(54) = .70$, $p < .001$, control: $r(56) = .81$, $p < .001$) were also larger than order recognition-cued recall correlations (imagery: $r(55) = .31$, $p = .020$, control: $r(54) = .37$, $p = .005$; Fisher test, imagery: $p < .001$, control: $p = .005$). Thus, consistent with Kato and Caplan (2017), order recognition exhibited a smaller correlation to cued recall accuracy than associative recognition.[6]

Importantly, Fisher tests between the control and imagery group OR-CR correlations were not significant pre- ($p = .85$) and post-instruction ($p = .70$), and AR-CR correlations pre- ($p = .57$) and post-instruction ($p = .15$), suggesting that imagery instructions did not affect the dependence of order or associative recognition on cued recall. This result does not support the hypothesis that imagery instructions help participants incorporate order. Instead, we have evidence for the

[5] We also found no support for the idea that significant pre-instruction PFT correlations were due to high PFT scorers spontaneously adopting imagery before being instructed to do so (page 50).

[6] When interpreting these results, one might consider the effect of testing pairs with cued recall before order recognition. Indeed, this was a major point addressed by Kato and Caplan (2017), who, in their second experiment, withheld half the pairs from cued recall testing, and in their third experiment, moved cued recall to the end of the session. In both cases they found that the order-cued recall relationship persisted, which we also found when analyzing testing effects in our own data-set, reported on page 47.

alternative hypothesis, that imagery does not change the formal characteristics of the association.[7]

Summary of experiment 1 Interactive imagery instructions increased cued recall accuracy and associative recognition d' above baseline, and compared to the control group. Imagery instructions did not improve order recognition, or change its relationship to cued recall. Both imagery vividness and skill did not predict the effectiveness of imagery instructions.

2.3 Experiment 2

The results of experiment 1 raised an additional question. Although interactive imagery failed to improve order recognition, if participants were given a specific way to incorporate order into their image, could that improve order recognition? We addressed this question by modifying the interactive imagery instruction in two ways (see Figure 2.1 for instructions). First, physically enacting verbal stimuli (e.g., hit the NAIL) improves benefits memory (enactment effects; cf. Allen, Waterman, Yang, & Jaroslawska, 2022; Engelkamp, 1991, 1995; Sivashankar & Fernandes, 2021), even when imagined (Allen et al., 2022; T. Yang et al., 2021). We hypothesized that imagining an actor–object relationship might not only exploit this benefit but also incorporate order into the image. Second, whereas the left–right axis is generally symmetric, gravity can break the symmetry; for example, a MOUSE on top of an ELEPHANT conjures a different meaning than the ELEPHANT on the MOUSE. We thus added two imagery instructions, where images were to comprise actor–object or top–bottom relationships, respectively.

Experiment 2 was pre-registered. All pre-registered analyses are reported. For analyses of the *within-subject* relationship of order/associative recognition to cued recall of pairs, see page 58.

2.3.1 Methods

Participants

Participants ($N = 433$) were recruited through Prolific (`www.prolific.co`), and compensated £6.50 for a 50-minute session. Participants were required to have English as their first language, be fluent in English, and have a Prolific approval rating above 70%. Our initial pre-registered exclusion criteria included failure to pass two attention checks, and/or exceeding a specified floor

[7] The *within-subject* analysis of the OR-CR relationship for experiment 1 and 2 are reported on page 50 and 58.

or ceiling threshold for recognition performance. Instead, we excluded participants who demonstrated clear evidence of disengagement, rather than exclude participants may have responded earnestly but performed extremely poorly or well: 13 were excluded because they re-wrote the presented probe in cued recall, suggesting they did not understand the task; three were excluded because they did not respond to any cued recall trial; seven were excluded because they responded to $< 10\%$ of recognition trials.

Groups Three main experimental groups were each divided into two sub-conditions: i) standard-imagery/associative recognition, ii) standard-imagery/order recognition, iii) actor-object/associative recognition, iv) actor-object/order recognition, v) top-bottom/associative recognition, vi) top-bottom/order recognition. Groups/sub-conditions were assigned with a random number generator function.

Materials and procedures

Materials and procedures were identical to experiment 1; however, with the following differences: (1) Experiment 2 was conducted online, with recruitment from `www.prolific.co`, hosted on `Pavlovia.org`. Groups were assigned with a random number generator. (2) The Paper Folding Task was omitted to save session time. (3) After the mid-session strategy instruction, participants were asked “Please explain back to us, in your own words, what we have asked you to do on the previous screen”. Short-answer responses were rated by two coders (KA and JT) blinded to group to quantify comprehension of instructions (corresponding on page 54).[8]. (4) After completing the VVIQ, participants rated, on a five-point scale, their frequency of incorporating mental imagery, interactivity, and order during study (page 53). (5) Participants answered a reversed-sense aphantasia question (see experiment 3 methods). Five aphantasic participants are presented as case studies in supplementary materials on page 58. (6) Two engagement checks were included; participants were presented a short message,“NOTE: Remember the number: X”, in the top-right corner of the screen, highlighted in blue, and against a grey foreground, once during the mid-session strategy instruction, and again, immediately after the VVIQ. Participants were asked to recall the number shortly after;however, two participants indicated their monitor cut off this number from the screen, thus, we applied different criteria, stated above. (7) Distractor trialswere held for a fixed 1000-ms period after the response was entered, regardless of response time. Additionally, there was a

[8] Note that in these analyses, we did not perform the chi-squared tests proposed in the pre-registration.

Table 2.2: Experiment 2: Included and excluded participants for each group and sub-condition. A total of 23 participants were excluded.

Group/Condition	Included	Excluded
Standard-imagery/Associative Recognition	73	2
Standard-imagery/Order Recognition	91	4
Actor-Object imagery/Associative Recognition	72	6
Actor-Object imagery/Order Recognition	68	3
Top-Bottom imagery/Associative Recognition	76	4
Top-Bottom imagery/Order Recognition	53	4

5000-ms maximum time-limit, and a blank 200-ms inter-trial interval. (8) Recognition trials were counterbalanced over all trials, including the practice list. However, there were two programming errors with associative recognition; i) a single recombined trial assigned to a list appeared as an intact trial, because could not exchange items with another pair.ii) random shuffling of recombined probes sometimes resulted in the original pairing. $N = 198$ participants had more intact probes than recombined probes, and of these participants, there was an average of nine extra intact trials. However, baseline associative recognition d' was comparable to experiment 1 (Figure 2.3), suggesting mean associative recognition performance was not sensitive to this design difference. (9) Recognition trials initially had a 15,000 ms time-limit. For d' calculations, rather than omit these trials from analyses outright, a correction was applied for each timed-out trial; if an intact trial was timed-out, 0.5 of an observation was added to hits and to misses. Likewise, if a recombined/reversed trial timed-out, 0.5 of an observation was added to false alarms and to correct rejections. In this way, timed-out trials pushed the overall d' to 0, where $d' = 0$ represents no memory, as if the participant was guessing. Thus, with this correction we assume that when a trial times-out, a participant has no knowledge, and would have guessed if given the opportunity. A total of 23 trials timed-out and were corrected in this manner. To remove the need for this estimation and obtain a response from each participant to each trial, time-limits were removed for recognition trials halfway through data-collection.

Distribution of VVIQ ratings VVIQ rating distributions were comparable to experiment 1 (Table 2.1).

2.3.2 Results and discussion

Cued recall A mixed ANOVA on cued recall accuracy (Figure 2.2) with design Group (standard-imagery, actor-object, top-bottom) × Instruction phase (pre-instruction, post-instruction) returned a significant main effect of Instruction phase, $F(1,430) = 71.13$, $MSE = 1.64$, $p < .001$, $\eta_p^2 = 0.14$, $BF_{\text{inclusion}} > 1000$. The main effect of Group was not significant, $F(2,430) = 1.15$, $MSE = 0.10$, $p = .32$, $\eta_p^2 = 0.005$, $BF_{\text{inclusion}} > 1000$, but had strong evidence in the Bayesian analysis[3].However, the Group × Instruction phase interaction was significant, $F(2,430) = 24.74$, $MSE = 0.57$, $p < .001$, $\eta_p^2 = 0.10$, $BF_{\text{inclusion}} > 1000$. Simple effects returned a supported null effect of Group pre-instruction ($p = .19$, $BF_{10} = 0.13$), but significant effect post-instruction ($p < .001$, $BF_{10} = 379.6$). Follow up t-tests on the post-instruction Group difference indicated a non-significant, supported null difference between the standard-imagery and actor-object imagery, $p = .19$, $BF_{10} = 0.29$. Additionally, cued recall accuracy was significantly lower in the top-bottom imagery compared to the standard-imagery ($p < .001$, $BF_{10} > 1000$), and actor-object ($p = .004$, $BF_{10} = 7.27$) imagery groups. Simple effects also returned a significant effect of Instruction phase for the actor-object, and standard-imagery group (both $p < .001$, $BF_{10} > 1000$), both of which increased in performance post-instruction, but a supported null difference for the top-bottom imagery group ($p = .60$, $BF_{10} = 0.11$). In sum, the actor-object imagery instructions matched the robust benefits of standard interactive imagery instructions for memory, but top-bottom instructions were ineffective.

Associative and order recognition Broadly speaking, the results for associative recognition paralleled those for cued recall; standard and actor-object imagery instructions were effective to improve performance and top-bottom instructions were ineffective. A mixed ANOVA on associative recognition d' (Figure 2.3), with design Group [3] × Instruction phase [2] returned significant main effects of Instruction phase, $F(1,195) = 21.38$, $MSE = 15.34$, $p < .001$, $\eta_p^2 = 0.10$, $BF_{\text{inclusion}} > 1000$, and significant Group × Instruction phase interaction, $F(2,195) = 7.56$, $MSE = 5.43$, $p < .001$, $\eta_p^2 = 0.07$, $BF_{\text{inclusion}} = 22.13$. Simple effects indicated that associative recognition performance increased post-instruction in both the actor-object group ($p = .003$, $BF_{10} = 9.65$)

and standard-imagery group ($p < .001$, $BF_{10} > 1000$), while the top-bottom group had a supported null difference between instruction phases ($p = .86$, $BF_{10} = 0.13$). Simple effects with the factor Group returned a supported null difference pre-instruction ($p = .34$, $BF_{10} = 0.16$), but a significant difference post-instruction ($p = .005$, $BF_{10} = 5.82$). Follow-up t-tests on the post-instruction group difference indicate that actor-object and standard-imagery had a supported null difference ($p = .84$, $BF_{10} = 0.21$), but both groups performed significantly better than the top-bottom group ($p = .017$, $BF_{10} = 3.75$ and $p = .003$, $BF_{10} = 9.86$ respectively).

Results for order recognition diverged from the other tasks. A mixed ANOVA on order recognition d' (Figure 2.3), with design Group [3] × Instruction phase [2] returned a significant main effect of Instruction phase, $F(1,232) = 12.89$, $MSE = 6.02$, $p < .001$, $\eta_p^2 = 0.053$, $BF_{\text{inclusion}} = 37.83$, indicating that order recognition d' improved in all three groups post-instruction. A significant improvement in order recognition somewhat diverged from null effects observed in experiment 1; however, the effect in all three groups was small in magnitude (d' post-minus-pre $\approx +0.25$, Figure 2.3), and post-instruction performance was in the range of values from experiment 1, suggesting the effect on order recognition was small in comparison to associative recognition. Importantly, both the main effect and interaction involving Group were supported null (both $p > .32$, $BF_{\text{inclusion}} < 0.3$), indicating that emphasizing order in the imagery instructions did not improve order recognition more than standard interactive imagery instructions.

The relationship between mental imagery vividness and the effectiveness of interactive imagery instructions VVIQ ratings had a supported null relationship to cued recall in three groups and instruction phases (all $p > .15$, $BF_{10} < 0.3$), replicating and extending findings from experiment 1 and 2. A single exception was found in the top-bottom imagery group pre-instruction, $r(54) = -.18$, $p = .03$, $BF_{10} = 1.09$, although a Bayesian correlation returned inconclusive evidence for this relationship (Tables 2.8–2.10). Correlations between VVIQ ratings and both order recognition, and associative recognition were non-significant, supported null effects (all $p > .36$, $BF_{10} < 0.31$). The failure to replicate the correlation between VVIQ and associative recognition in experiment 1 suggests that this finding is not particularly robust and will not be discussed further. Thus, vividness ratings in the VVIQ could not explain the advantage of standard-imagery instructions, nor memory performance under any imagery instruction variant.

The relationship of order recognition to cued-recall Due to low trial counts for recombined trials (see Methods), the associative recognition measures are noisy and should be interpreted with caution. However, with maximal power by collapsing across groups (Figure 2.19, Table 2.3), the OR-CR correlation was significantly lower than the AR-CR correlation, both pre- and post-instruction ($p = .047$, $p = .0034$ respectively, Fisher tests), replicating experiment 1 and Kato and Caplan (2017). Next, we asked if, for any instruction, the OR-CR correlation changed from pre- to post-instruction. These comparisons were non-significant for top-bottom ($p = .71$, Fisher test) and actor-object group ($p = .63$), but there was a significant decrease post-instruction for the standard-imagery group ($p = .034$). This pre- versus post-instruction difference in the standard-imagery group was largely driven by a single outlier (Figure 2.19) who performed extremely poorly in cued recall, but extremely well in order recognition. When removed, the comparison was non-significant ($p = .14$).

Summary of experiment 2 Standard interactive imagery and actor-object imagery instructions boosted cued recall and associative recognition above baseline, and compared to the top-bottom imagery instructions. Surprisingly, both imagery instructions that emphasized order had a negligible effect on order recognition, and did not affect its relationship to cued recall. Replicating experiment 1, imagery vividness did not predict the effectiveness of imagery instructions.

2.4 Experiment 3

Experiment 1 suggested the large benefit to cued recall of interactive imagery has little to do with subjective detail or objective visual imagery skill. In experiment 3, we recruited aphantasics, who self-report an inability to form visual imagery, and non-aphantasics, to do cued recall, VVIQ and PFT as in experiment 1. If the presence of visual images is required for interactive imagery, then aphantasics should show substantially less benefit from imagery instructions than non-aphantasics.

2.4.1 Methods

Participants

Just as in experiment 1, participants ($N = 122$) were enrolled in an introductory psychology class at the University of Alberta, and recruitment had the same basic restrictions. Participants who

Table 2.3: Experiment 2: Correlations between log-odds cued recall accuracy and both order and associative recognition collapsed across participants, and separated into groups.

	Pre-instruction		Post-instruction	
	r	p	r	p
All Participants/Associative Recognition	.67	$< .001$	.66	$< .001$
All Participants/Order Recognition	.54	$< .001$	.47	$< .001$
All Participants Fisher test (Order versus Associative)	$z = 1.99$, $p = .047$		$z = 2.93$, $p = .003$	
Standard-imagery/Associative Recognition	.64	$< .001$	.72	$< .001$
Standard-imagery/Order Recognition	.58	$< .001$	.32	.0017
Standard-imagery Fisher test (Order versus Associative)	$z = 0.59$, $p = .55$		$z = 3.62$, $p = .0003$	
Actor-Object/Associative Recognition	.70	$< .001$	.66	$< .001$
Actor-Object/Order Recognition	.44	$< .001$	.50	$< .001$
Actor-Object Fisher test (Order versus Associative)	$z = 2.18$, $p = .030$		$z = 1.34$, $p = .18$	
Top-Bottom/Associative Recognition	.67	$< .001$	.61	$< .001$
Top-Bottom/Order Recognition	.64	$< .001$	.68	$< .001$
Top-Bottom Fisher test (Order versus Associative)	$z = 0.29$, $p = .77$		$z = 0.64$, $p = .53$	

had enrolled in experiment 1 were not permitted to participate in this study. Four participants were excluded from analyses because they accessed the online link and completed the experiment twice; both sessions were excluded. One participant was excluded for providing no cued recall or math distractor responses.

Recruitment Before the experimental session, potential aphantasics and non-aphantasics were identified via online mass-testing questionnaires administered to University of Alberta introductory psychology students at the beginning of the Fall 2020 ($N = 2357$) and Winter 2021 ($N = 1975$) semesters. Along with many other items that were part of different studies, questionnaire participants responded yes/no to “Are you able to form mental images (i.e., pictures) in your mind’s eye?”.

Recruitment for experiment 3 was conducted after the Winter 2021 questionnaire was administered, and was restricted to participants who responded to this question in *either* the Fall or Winter questionnaire. We note here that filling out a mass questionnaire did not guarantee that a student signed-up for our experiment. Participants could only sign up if they had answered the aphantasia question in the mass-testing. A different project code was visible to those who answered yes and no, respectively, to roughly equate recruitment rates. However, we further classified the 122 who participated with the additional in-session, reversed-sense aphantasia question.

Aphantasia classification We classified aphantasia in these 122 participants based on three different criteria, which we call “consistent”, “moderate” and “extreme” aphantasics, respectively.

The first criterion was based on consistent response to the yes/no aphantasia question. Participants who consistently indicated being unable to form mental images in mass-testing and in-session, were classified as “consistent aphantasic” ($N = 25$). Those who consistently indicated the opposite were “consistent non-aphantasic” ($N = 34$). Those who were inconsistent in their responses to this question formed a third “inconsistent-responder” group ($N = 64$). Because inconsistent responders changed their answers across testing sessions, we were hesitant to classify them as either aphantasic or non-aphantasic, as they might have been unsure of their status. Additionally, because the recruitment question was embedded within a much longer questionnaire this raised the possibility that individuals would not respond conscientiously to each questionnaire item. This provided more reason for classifying aphantasia based on multiple responses.

To be more selective, we also applied more conservative second and third criteria from Zeman et al. (2020). Of the “consistent aphantasics,” participants rating 73–79 (maximum 80) VVIQ in-session were considered “moderate” aphantasics ($N = 7$), while ratings of 80/80 were considered “extreme” aphantasics ($N = 3$). VVIQ criterion aphantasic participants are reported as case studies (Table 2.4).

A strength of our procedure was that our experimental session was separated by days or weeks from the Winter mass-testing questionnaire. The in-session reversed-sense aphantasia question and VVIQ were at the end of the session. We thought this should make the constructs of aphantasia and even visual imagery less front-of-mind for participants than in previous aphantasia studies.

Mass questionnaire aphantasia prevalence rates Next, we applied our three aphantasia classification criteria to mass questionnaire data to provide an estimate of the prevalence of aphantasia in our student population. Note that the following numbers are based *solely* on mass questionnaire data and and not on the sub-sample tested with memory tasks in experiment 3.

We identified 772 participants who answered the aphantasia question in both the Fall and Winter mass testing sessions. Of these participants, 30 indicated being unable to form mental images in both sessions (3.9%). This approached Faw’s (2009) previously estimated rate of 2–3%.

Our conservative aphantasia classification criteria based on VVIQ cutoffs were identical to Zeman et al. (2020), who observed the rate of moderate aphantasia ($73-79/80$) and extreme aphantasia ($80/80$) to be 2.6% and 0.7% in their mass-testing questionnaire. First, of the $N = 2000$ who completed the VVIQ in the Fall 2020 mass-testing, 23 (0.9%) and 9 (0.4%) met these VVIQ cutoffs respectively. Next, of the 1975 participants who responded to the VVIQ in Winter 2021 mass testing questionnaire, 43 (2.2%) and 26 (1.3%) participants met the moderate and extreme VVIQ cutoffs respectively. In sum, the prevalence rates that were derived from the Fall 2020 questionnaire were considerably lower than previous observations, while the rates that were derived from the Winter 2021 questionnaire were closer to Zeman et al. (2020). The extreme cutoff appears far more highly selected than prior aphantasic samples.

Materials and procedures

Materials and procedures were identical to experiment 1 except: (1) This experiment was conducted completely online, on `Pavlovia.org`. The experiment was created using the PsychoPy

Builder interface (Peirce et al., 2019) and translated to a PsychoJS experiment (Bridges, Pitiot, MacAskill, & Pierce, 2020). As in experiment 1, recruitment was conducted through the University of Alberta psychology research participation pool, but participants completed the experiment on their personal devices. (2) All participants were instructed to use interactive imagery half-way through the session (no control group) (3) Recognition tasks were omitted; pairs were only tested with cued recall. (4) To use the additional testing time freed up from the recognition tasks, participants studied 10 lists (cf. eight in experiment 1). (5) The PFT was re-added to the design, and administered after the VVIQ just like in experiment 1. (6) After the PFT, participants answered a single free-form question about their strategy-use question. (7) Cued recall direction (forward versus backward) was counterbalanced over all trials, including the practice list. (8) After the strategy-use question (i.e., at the end of the session) a reversed-sense version of the aphantasia recruitment question was administered: "Are you unable to form mental images (i.e., pictures) in your mind's eye?". (9) Distractor trials were identical to experiment 2, except that immediately after the response was entered, the screen was held for 2000-ms fixed period (versus the 1000-ms fixed period in experiment 2).

VVIQ test-retest reliability We analyzed test-retest reliability of the VVIQ between mass questionnaires and the in-session administration, reported on page 63.

Analysis of gender and interactive imagery effects We obtained data on self-reported gender for participants in experiment 3. These are reported on page 62.

Free-form strategy self report After the PFT, participants were asked to "describe how you studied the word pairs, whether or not that included the use of visual imagery as instructed, in a short one or two sentence response." These responses were rated by two coders, blinded to condition, for two measures of interest. Firstly, rated either 1) response includes imagery, 2) response explicitly excludes imagery, 3) response leaves open the possibility of imagery but was not explicit. Second, rated for whether it referred to interactivity or connection between words (yes/no). Analyses incorporating these ratings are reported on page 61.

2.4.2 Results and discussion

Of 122 participants, 25 were consistent aphantasics, 34 were consistent non-aphantasic and 63 were inconsistent responders.

Self-reported vividness Supporting the validity of our yes/no aphantasia self-identification question, consistent aphantasic responders scored significantly higher (lower vividness) than the non-aphantasic group ($p < .001$, Mann-Whitney U test[9]) and the inconsistent responder group ($p < .001$) on the VVIQ, where higher scores indicate lower vividness. The difference between inconsistent responders and consistent non-aphantasic responders nearly reached significance ($p = .07$). Additionally, the average VVIQ rating for consistent aphantasic responders was well above values in experiments 1 and 2 (Table 2.1). Visual inspection reveals a number of characteristics of the VVIQ responses. First, the inconsistent responders contained participants who exhibited both extremely high and extremely low vividness. Second, a sizeable number of consistent aphantasics nonetheless reported moderate amounts of vividness in the VVIQ, with ratings within the middle of the VVIQ distribution for consistent non-aphantasics. We do not think that participants are simultaneously reporting an inability to form images (aphantasia question) while reporting vivid mental images (VVIQ). Instead, consistent aphantasics who rated high vividness might have either responded carelessly, or interpreted vividness in terms of the amount of detail within a non-visual representation.

Cued recall A mixed ANOVA on cued recall accuracy (Figure 2.2), with design Group (consistent aphantasic, inconsistent responders, consistent non-aphantasics) × Instruction phase (pre-instruction, post-instruction), returned a significant main effect of Instruction phase, $F(1,119) = 91.02$, $MSE = 1.59$, $p < .001$, $n_p^2 = 0.43$, $BF_{\text{inclusion}} > 1000$. However, Group, and Group × Instruction phase, were supported null effects (all $p > .5$, $BF_{\text{inclusion}} < 0.3$), indicating that aphantasia status did not influence the benefit of interactive imagery instructions. Additionally, the cued recall accuracy achieved after the imagery instruction in each group was comparable to the imagery group from experiment 1 ($\approx 60\%$), suggesting that the imagery manipulation was successful, and all three groups from experiment 3 would presumably have scored higher than a control group, had

[9] We tested group differences with non-parametric tests due to the skewed vividness rating distribution in the consistent aphantasia group (Figure 2.4).

it been included.

Paper-folding task A one-way ANOVA on PFT accuracy with Group[3] returned non-significant, supported null effect ($p = .52$, $BF_{\text{inclusion}} < 0.3$), and likewise for PFT response time ($p = .83$, $BF_{\text{inclusion}} < 0.3$). Thus, aphantasic participants did not exhibit worse visuospatial skill, measured objectively, and achieved comparable scores to participants in other experiments (Table 2.1). These results suggest that the PFT may be added to a class of visuospatial tasks for which aphantasics are fully competent (Zeman et al., 2020), such as mental rotation (Shepard & Metzler, 1973), and the Brooks' matrix spatial task (Brooks, 1967), which we revisit in the general discussion.

The relationship among mental imagery skill, vividness, and the effectiveness of interactive imagery instructions First, including all participants, VVIQ ratings had a supported null correlation with cued recall accuracy (both $p > .39$, $BF_{10} < 0.30$), and both PFT accuracy and response times had a positive correlation to cued recall accuracy in both instruction phases (Table 2.13), replicating experiment 1, and with broader coverage of the range of VVIQ values.

Next, we asked whether variability within each group of participants might show different effects. With correlations computed separately for consistent aphantasics, consistent non-aphantasics and inconsistent responders, VVIQ ratings again had a supported null relationship to cued recall accuracy in both instruction phases and all groups ($p > .29$, $BF_{10} < 0.36$), except for inconsistent responders in the pre-instruction phase, $r(61) = -.27$, $p = .03$, $BF_{10} = 1.42$, although the Bayesian correlation was inconclusive. Importantly, VVIQ ratings did not determine the effectiveness of the interactive imagery within the group of consistent aphantasics.

PFT accuracy positively correlated with cued recall accuracy for all three groups and in both instruction phases, and PFT response time had significant positive correlations with cued recall accuracy in both the pre- and post-instruction phases. Thus, skill on this visuospatial task did not predict the effectiveness of interactive imagery even within the consistent aphantasic group.

More conservative criteria for aphantasia Next, we applied increasingly conservative criteria for classification of aphantasics, as described in the Methods. Given the low numbers, these should be interpreted as multiple case studies. Our goal was to check if applying more strict classification criteria would show hints of increased group differences, even while reducing statistical power.

Inconsistent with this, three one-way ANOVAs, with factor Group (VVIQ criterion consistent aphantasics, non-VVIQ criterion consistent aphantasics, inconsistent responders, consistent non-aphantasics) on PFT accuracy, PFT response time and Change in Accuracy returned favoured null effects of Group (all $p > .57$, $BF_{\text{inclusion}} < 0.3$). Five of the 10 VVIQ criterion participants reported, unprompted, difficulty forming visual images. Eight exhibited at least a 10% increase in cued recall following the imagery instruction, with four increasing by 22.5% or more.

Eight participants explicitly reported the use of alternative strategies. It was unclear if participant 1 was referring to mental imagery or not, but described some difficulty with imagining and resorting to "memory of thinking about it". Two participants (7 and 9) reported rote repetition, known to be a poor associative strategy (Bower & Winzenz, 1970), yet still increased substantially (+22.5% and +15%). Two participants did not benefit from the imagery instruction; participant 3 exhibited a small negative change (−2.5%), participant 5 exhibited a substantial reduction (−25%) in performance and, interestingly, was the only VVIQ criterion aphantasic who reported trying to implement imagery instructions, suggesting that strict adherence to the imagery instructions may not be beneficial to aphantasics.

Our extreme aphantasics, participants 4, 6, and 7, are of particular interest. Each reported no vividness, were perfectly consistent across multiple administrations of the aphantasia question, and described using non-imagery strategies, consistent with their complete lack of mental imagery. All three benefited from the imagery instruction (+10%,+10%, and +22.5%).

In sum, the reduction in sample size was not offset by any hint of an emerging deficit of aphantasics to respond to interactive imagery instructions, converging with our other evidence against the centrality of visual imagery for interactive imagery instructions.

2.5 General Discussion

We replicated the positive effect of interactive imagery instructions on cued recall (Bower & Winzenz, 1970; Bower, 1970a; Paivio, 1969; Paivio & Yuille, 1969; Paivio & Foth, 1970; Richardson, 1985, 1998) compared to control instructions (experiment 1), compared to the no-instruction baseline (all experiments), and compared to the "top-bottom" variant of standard interactive imagery instructions (experiment 2). Correlations between characteristics of a participant's visual imagery (individual differences in visuospatial skill and vividness) and the effectiveness of inter-

Table 2.4: Experiment 3: Change in cued recall accuracy, strategy self-report, VVIQ rating, PFT accuracy and response times for “consistent aphantasics” who scored higher than 73 on the VVIQ. Responses from extreme aphantasic participants who rated 80/80 on the VVIQ are in bold.

Participant	Change in Accuracy.	Strategy self report	VVIQ (out of 80)	PFT accuracy (out of 20)	PFT response time (seconds)
1	+22.5%	“I chose to use visual imagery or the memory of thinking about it since I have trouble imagining things in my mind.”	77	16	8.84
2	+10%	“I did attempt to do as asked for some of the pairs but I also tried to use short phrases to remember alongside the imagery.”	76	10	27.77
3	−2.5%	“I tried to remember any word combinations that stood out based on if they made sense together or not or if the words presented were relevant to me.”	78	7	6.02
4	+10%	**“I cannot really picture things so I just said the words out loud and tried to create jokes that included both words as they came up.”**	80	15	19.65
5	−25%	“Initially I was saying associations out loud and that worked well, then with the imagery it was hard because I have a hard time invisioning things quickly and alot of the images would have multiple aspects so I would get confused on what I was meaning to associate.”	76	12	17.38
6	+10%	**“I tried to find a connection between the two words so I can remember them better.”**	80	16	19.25
7	+22.5%	**“I said the words aloud as they appeared in pairs and didnt do the visualisation thing.”**	80	19	24.61
8	+25%	“In the beginning I was trying to memorize them just by saying them but when you told me to memorize them by thinking of an image with them I would think of a scenario where the two words would go together for example ice cream and mistake would be dropping ice cream.”	76	4	14.20
9	+15%	“I cant picture anything in my mind so I couldnt do that, I just kept repeating the words as many times before they disappeared.”	77	12	13.55
10	+32.5%	“I attempted to use visual imagery but I cant get a visual imagine in my mind so I just thought of short scenarios of the two words merged together.”	78	10	13.69

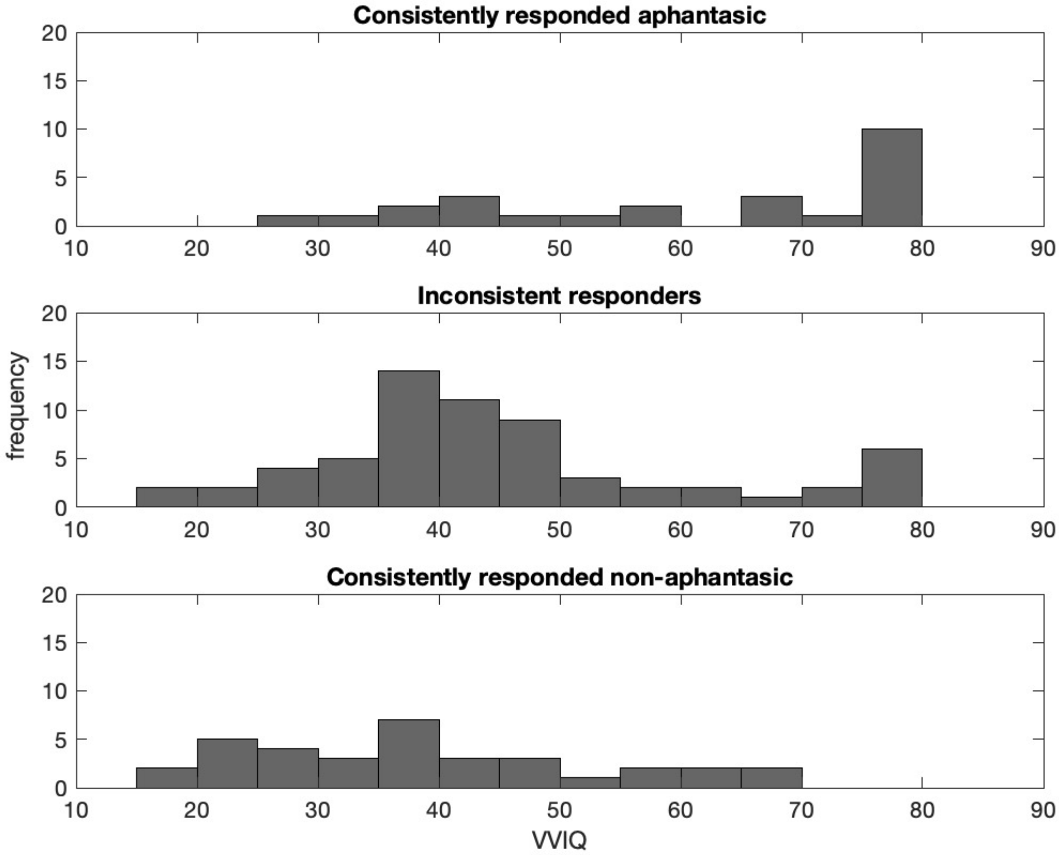


Figure 2.4: Experiment 3: Distributions of VVIQ responses for experimental group from experiment 3. Note, lower scores indicate higher vividness.

active imagery produced supported null effects.[10] Furthermore, aphantasics showed no trace of impairment despite their self-diagnosed inability to form visual imagery (experiment 3). Thus, we found no support for the hypothesis that visual images are necessary for interactive imagery benefits, raising the possibility of alternative explanations.

Curiously, order recognition was not improved by interactive imagery (experiment 1), nor even instructions incorporating order into the image (experiment 2). Whatever additional detail/information is afforded by interactive imagery instructions evidently does not provide order. Moreover, the relationship between order recognition and cued recall was not influenced by instruction. These results argue against the hypothesis that imagery strategies result in formally different association memories that contain more order. Instead, our results were more consistent with the alternative hypothesis that imagery produces associations that are qualitatively the same as non-imagery conditions.

Subjective vividness does not explain imagery-instruction benefits to cued recall In all three experiments, subjective vividness of mental imagery (VVIQ rating) did not explain the effectiveness of interactive imagery for cued recall. This was reinforced in experiment 3, where aphantasics (high VVIQ) benefited from interactive imagery instructions as much as others (Figure 2.2). All VVIQ-criterion aphantasics that benefited post-instruction reported either solely using non-imagery strategies or a combination of imagery and non-imagery strategies, but evidently with no consequence for their benefit from interactive imagery instructions. Even three participants who reported exactly no vividness benefited from imagery instructions while reporting using imagery-free strategies. This seems consistent with the observation that congenitally blind participants can effectively apply the Method of Loci, which is typically described as heavily dependent upon visual imagery (de Beni & Cornoldi, 1985), and with null correlations of the VVIQ with this strategy (Kliegl et al., 1990; Kluger et al., 2022).

Although the VVIQ has been widely used to assess subjective imagery vividness (Marks, 1973), and is a primary way to classify aphantasia (Zeman et al., 2015), there have been specific critiques about its content validity that may be important to consider (McKelvie, 1995; Pylyshyn, 2002). McKelvie (1995) suggested the VVIQ may not capture important dimensions of imagery experience, such as the distinction between imagery vividness and generation. Future studies

[10] Additionally, correlations between post-minus-pre instruction memory performance and our visual imagery measures produced supported null effects (page 64)

should focus on qualities of visual imagery experience that the VVIQ may not adequately capture, like imagery generation.

Objective imagery skill does not relate to interactive imagery PFT accuracy did not predict the effectiveness of the interactive imagery instructions, but covaried with performance even before strategy instructions were given (experiments 1 and 2). Although this does not rule out the PFT as a measure of other memory processes like working memory or visuospatial ability, it weakens the argument that imagery skill determines success with interactive imagery instructions.

Interestingly, there was a supported null difference between PFT performance in aphantasics and non-aphantasics in experiment 3, which may place the PFT in a class of visuospatial tasks that aphantasics perform without any clear deficits (Zeman et al., 2010). Both Zeman et al. (2010) and Bainbridge et al. (2021) suggested that aphantasics use symbolic/verbal strategies for visuospatial tasks. Thus, the cognitive processes required for this task may not necessarily depend on visual images, which suggests a dissociation between conscious mental imagery experience and the cognitive processes engaged when solving complex visuospatial problems. Furthermore, because the PFT could not explain the benefits of interactive imagery, its intact status in aphantasics cannot explain why aphantasics showed virtually no reduced benefit from these instructions.

Validity of aphantasia-status classified by self-report Our three criteria for classifying aphantasia in experiment 3 (multiple consistent responses to the aphantasia recruitment question, and two VVIQ cutoffs), produced prevalence rates that approached the estimates in previous studies (see methods), suggesting that methods of classifying aphantasia in experiment 3 aligned well with previous aphantasia studies. Despite this, there are broader critiques of classifying aphantasia by self-report. For example, de Vito and Bartolomeo (2016) suggested aphantasics may underestimate a latent ability to form mental images. Perceived absence of mental imagery experience may then be due to poor/altered meta-cognition rather than fundamental differences in cognitive representations. However, even if aphantasia is due to an inaccurate sense of one's own imagery ability, our findings still show that this kind of imagery self-efficacy is immaterial to memory-success following interactive imagery instructions, again problematic for the hypothesis that interactive imagery acts through the formed image, itself.

Interactive-imagery effects without visual imagery Our findings challenge the notion that visual imagery, in any literal sense, is essential for the benefit to cued recall of interactive imagery instructions. In other words, the subjective experience of mental imagery is experienced by those who are able, but is not required for later memory benefits. This resonates with Pylyshyn's (2002) argument that the experience of mental imagery may be epiphenomenal, and not necessarily causal.

A similar story is emerging from recent research on word concreteness/imageability effects. High-imageability words are recalled better low-imageability words (Paivio, 1969). Hockley (1994) found better associative recognition for higher concreteness word pairs. Paivio and colleagues explained concreteness as providing participants the greater availability to construct visual image mediators for concrete/imageable than abstract/low-imageable words, confirmed by findings of more frequent self-reported use of imagery strategies during the study of high imageability word pairs (Paivio, Smythe, & Yuille, 1968; Paivio & Yuille, 1969). Thus, the historical understanding of the concreteness/imageability effects is functionally linked to visual imagery-related strategies like interactive imagery.

However, behavioural and neuroimaging findings have challenged the idea that concreteness effects can be explained via visual imagery. Westbury et al. (2013) and Westbury, Cribben, and Cummine (2016) showed that concreteness effects on lexical decision could be explained by non-imagery factors like size/density of a word's context and its emotional associations (see Fiebach & Friederici, 2004, and see Cox, Hemmer, Aue, & Criss, 2018 who found semantic diversity, alongside concreteness, to be a strong predictor of memory performance). In neuroimaging studies, one can look for memory-related activity in brain regions that are involved in mental imagery, such as posterior visual-processing regions and right-lateralized activity. However, Caplan and Madan (2016) found no brain activity reminiscent of visual imagery explaining word-imageability effects on cued recall (see also Klaver et al., 2005). Rather, higher imageability was associated with more hippocampal activity (somewhat left-dominant), which in turn, apparently increased memory. Similarly, Duncan, Tompary, and Davachi (2014) found that functional connectivity between hippocampus and ventral tegmental area during interactive-imagery instructions predicted retrieval success, regions that are not specialized for imagery.

An alternative explanation of interactive imagery effects Vincente and Wang (1998) emphasized the idea that expert-memory effects depend on participants engaging with stimuli in a manner that is relevant to their expert domain. Extrapolating to non-expert domains, perhaps interactive-imagery acts primarily by inspiring participants to engage with word pairs in a manner that leads to this kind of meaningful or deep processing. But what is the nature of this deeper processing, and how does it improve memory? Some hints may be gleaned from experiment 2. Standard-imagery and actor-object imagery both resulted in benefits to memory. Given the high similarity between the examples given for both instructions, both instructions may have engaged the same mechanisms, perhaps revealing some role of motor imagery (Allen et al., 2022; T. Yang et al., 2021) in interactive imagery effects. In contrast, top-bottom instructions which ask participants to imagine a spatially organized image including both words, and do not explicitly refer to the words interacting, did not change cued recall or associative recognition from baseline. Top-bottom imagery may be difficult to implement, especially for certain word pairs. For example, it is easier to conceptualize a spatially organized image of APPLE DRAGON, compared to ASPECT LEVEL (both of which were possible pairings in our study); however, this challenge would also exist with standard and actor-object strategies (concreteness effects; cf. Hockley, 1994; Paivio, 1969). Alternatively, top-bottom instructions may miss a key component— explicit instructions to conceptualize an interactive, functional relationship between the items. Top-bottom imagery may resemble explicitly non-interactive "separation-imagery" instructions, where participants are asked to form mental images of each word in isolation, which does not improve association-memory (Bower, 1970a; Dempster & Rohwer, 1974; Hockley & Cristi, 1996).

In contrast, by leading participants to think about an interactive relationship between words, effective associative strategies like interactive imagery may facilitate encoding of additional item features that are pair-unique. To illustrate how this may occur, consider an associative recognition task for the pairs APPLE TEACHER and TABLE OVEN. An image (or non-visual analogue) of a TEACHER with an APPLE (intact, here) may generate a stereotypical image of a crisp, red apple on a teacher's desk, whereas an image of an OVEN with a APPLE (recombined, here) might bring to mind baked apples. The more a participant focuses on how the words might interact, the more detailed and pair-specific the stored representations might be (see the modelling work of Caplan, Chakravarty, & Dittmann, 2021, Cox & Criss, 2017, 2020, and Benjamin, 2010). For example, Cox and Criss (2020) showed how similarity can cause the representations of two items to become correlated, by drawing attention to their common features. One intriguing possibility is that interactive imagery amplifies this very same effect by drawing the participant's attention to shared features.

Supporting encoding of more detailed item representations, item recognition improves alongside associative memory performance, when comparing interactive imagery to rote repetition (Dempster & Rohwer, 1974; Hockley & Cristi, 1996).[11] Such a mechanism could conceivably occur without visual imagery. This is consistent with findings that verbally mediated strategies for association-memory (e.g., form a sentence including both words) are nearly as effective (Dunlosky et al., 2005; Hockley & Cristi, 1996).

Interactive imagery instructions do not change model-relevant characteristics of the association Largely replicating and extending the boundary conditions of Kato and Caplan (2017), order recognition significantly correlated with cued recall accuracy, but significantly weaker than the correlation between associative recognition and cued recall (Figures 2.15, 2.19, and 2.20). Despite large effects on association-memory, imagery instructions did not modulate these findings (Figures 2.15, 2.19, and 2.20). Whatever additional detail/information is afforded by imagery instructions does not improve memory for order. An interesting possibility here is that order and associative information are somehow represented differently in memory, explaining why manipulations of association-memory do not affect memory for order. Cox and Criss (2020) suggested order could be represented by item features distinct from associative features. In any case, our findings indicate that challenges to perfect-order models, which predict a perfect relationship between order recognition and cued recall, and order-absent models, which predict no relationship, are not particular to uninstructed participants, but generalize to several instructed strategies. This increases the need for models that can accommodate moderate-level order within associations.

2.6 Conclusion

Interactive-imagery instructions improve associative memory without requiring vividness, visual-imagery skill, nor even the subjective sense that one can create visual imagery. The instruction may instead lead participants to conceptualize elaborate, interactive relationships, leading to storage of more distinctive features. Finally, whatever additional detail aids associative memory does not

[11] Both Hockley and Cristi (1996) and Dempster and Rohwer (1974) also found that separation imagery improved item recognition, suggesting that interactivity is not *required* to encode more detailed item representations. However, the additional item features granted by non-interactive strategies would likely not be pair-specific, which may explain the lack of effects on associative memory.

provide order.

2.7 Supplementary Materials

2.7.1 Experiment 1

Correlations between visual imagery measures and memory performance

Tables 2.5–2.7 report each correlation between visual imagery measures (PFT and VVIQ) and performance in cued recall, associative recognition and order recognition tasks.

Scatter plots of visual imagery measures versus memory performance

Figures 2.5–2.13 are scatter-plots corresponding to each correlation reported in Tables 2.5–2.7.

The effect of cued recall direction on order recognition performance

Because Kato and Caplan (2017) found that cued recall in the forward direction increased order recognition of a pair whereas cued recall in the backward direction reduced order recognition, the following analyses test if cued recall direction (forward versus backward) affected order recognition and its relationship to cued recall in our data.

First, a mixed ANOVA on mean order recognition d' (Figure 2.14), with design Group (imagery-order recognition, control-order recognition) $\times$ Instruction phase (pre-instruction, post-instruction) $\times$ Cued recall direction (forward, backward) returned a significant main effect of cued recall direction, $F(1,111) = 68.0$, $MSE = 34.81$, $p < .001$, $\eta_p^2 = 0.38$, $BF_{\text{inclusion}} > 1000$, replicating the finding that order recognition was better overall for pairs tested with forward cued recall (Kato & Caplan, 2017). Group $\times$ Instruction phase nearly reached significance, $F(1,111) = 3.73$, $MSE = 2.06$, $p = .056$, $\eta_p^2 = 0.032$, $BF_{\text{inclusion}} = 0.13$, although the Bayesian analysis returned supported null evidence (see experiment 1 in main text for an analysis of this interaction, which indicated control participants became worse at order recognition as the experiment progressed, while imagery participants did not change). All other effects were supported null (all $p > .12$, $BF_{\text{inclusion}} < 0.3$). In sum, although order recognition was better for pairs tested with forward cued recall, cued recall direction did not change the null effect of imagery instructions on order recognition performance.

Next, to test if direction of cued recall affected the relationship between order recognition and cued recall, we also calculated between-subject correlations between log-odds cued recall accuracy to both order and associative recognition d', split by direction of the cued-recall test. Scatter plots of all of these correlations are plotted in Figure 2.16 for the control group, and in Figure 2.17 for the imagery group, and reported in Table 2.12.

In brief, beyond the overall difference in d', the pattern of results for pairs tested forward was quite similar to the pattern for pairs tested backward. With only one exception, the correlation between order recognition and log-odds cued recall was significantly smaller than the control correlation (associative recognition-log-odds cued recall) in all groups and instruction phases, regardless of whether recognition involved pairs tested prior with backward and forward recall. In sum, cued recall direction did not seem to affect model-relevant patterns that indicate order recognition has a mid-range relationship to cued recall.

Self-report on strategy use

Halfway through data collection we included a section at the end of experiment 1 (i.e., after the PFT) where both control and imagery groups were given an opportunity to rate how often they used interactive imagery in both phases of the experiment 1 (e.g., 1: never, 2: sometimes, 3: mostly, 4: always), and provide a free-form response about their strategy use. Because the control group had not encountered interactive imagery instructions, the strategy was described to them before they provided ratings or responses. The imagery group was reminded of the strategy before they provided ratings and responses. To test whether self-report on imagery strategy use had any relationship with objective task performance, we examined the relationship between both pre- and post-instruction ratings and the change in cued recall accuracy (e.g., accuracy post minus pre) below.

Pre-instruction An ANOVA on Group (imagery, control) × Pre-instruction Imagery rating (never, sometimes, mostly, always) returned significant main effects of Group, $F(1,119) = 4.38$, $MSE = 0.18$, $p = .039$, $\eta_p^2 = 0.035$, $BF_{\text{inclusion}} > 1000$, Pre-instruction Imagery rating, $F(3,119) = 9.54$, $MSE = 0.40$, $p < .001$, $\eta_p^2 = 0.19$, $BF_{\text{inclusion}} > 1000$, and interaction Group × Pre-instruction Imagery rating, $F(3,119) = 4.19$, $MSE = 0.18$, $p = .007$, $\eta_p^2 = 0.095$, $BF_{\text{inclusion}} = 31.10$. Tukey t-tests indicated that *imagery* participants who rated “never” exhibited a significantly larger increase in cued recall accuracy than *imagery* participants who rated “sometimes” ($p_{\text{tukey}} = .005$), ”mostly” ($p_{\text{tukey}} < .001$), and ”always” ($p_{\text{tukey}} = .018$), and *control* participants who rated “never” ($p_{\text{tukey}} = .008$), “sometimes” ($p_{\text{tukey}} < .001$), ”mostly” ($p_{\text{tukey}} < .001$), and ”always” ($p_{\text{tukey}} < .001$). This indicates that participants who reported never using interactive imagery pre-instruction received the most benefit. All other pre-instruction ratings did not differ significantly from each other in the imagery group (all $p_{\text{tukey}} > .12$). In sum, participants who reported no spontaneous use of interactive imagery pre-instruction received the most benefits from imagery instructions.

Post-instruction An ANOVA on Group [2] × Post-instruction Imagery rating [4] returned a significant main effect of Post-instruction Imagery rating, $F(3,119) = 5.48$, $MSE = 0.26$, $p = .001$, $\eta_p^2 = 0.12$, $BF_{\text{inclusion}} = 185.91$. The effects of Group, ($p = .078$, $BF_{\text{inclusion}} = 87.18$), and the interaction Group × Post-instruction Imagery rating, ($p = .69$, $BF_{\text{inclusion}} = 0.79$) were not significant, although $BF_{\text{inclusion}}$ values indicated strong evidence for Group. Tukey t-tests indicated that, irrespective of group, participants who rated “always” exhibited significantly larger increases in cued recall accuracy than participants who rated “sometimes” ($p_{\text{tukey}} < .001$), but were not significantly different than participants who rated “never”, or “mostly” (both $p_{\text{tukey}} > .19$), suggesting a positive effect of compliance with instructions. Additionally, there was a trend towards participants who rated “mostly”, exhibiting more benefits than participants who rated “sometimes”, although this difference fell just short of significance ($p_{\text{tukey}} = .061$). Thus, as would be expected, participants who self-reported “always” in the post-instruction phase exhibited the largest increases in cued recall accuracy, although the effect was not large enough to reach significance over participants who indicated “never”.

Imagery vividness/ability and the spontaneous use of interactive imagery. We considered the possibility that participants high in imagery vividness (VVIQ) or ability (PFT) might have been more likely to have adopted imagery spontaneously pre-instruction, which would complicate the interpretation of several of our results.

Participants who rated that they never used imagery pre-instruction exhibited a larger imagery benefit to cued recall accuracy, suggesting participants had reliable retrospective insight into their strategy use during the experiment. We were motivated to look for evidence that participants who provided different ratings also had different PFT accuracy, PFT response times, and/or VVIQ ratings. An ANOVA on PFT accuracy with one factor Pre-instruction Imagery rating (never, sometimes, mostly, always) returned a supported null-effect ($p = .99$, $BF_{\text{inclusion}} < 0.3$), and likewise for PFT response times ($p = .36$, $BF_{\text{inclusion}} < 0.3$), or VVIQ ratings ($p = .21$, $BF_{\text{inclusion}} = 0.398$), arguing against the idea that participants with high imagery skill/vividness were more likely to spontaneously use imagery as a strategy.

The effectiveness of interactive imagery instructions based on pre-instruction performance.

To check if imagery instructions would be more effective for participants with poor baseline performance. Indeed, correlations between pre-instruction memory performance and post-minus-pre instruction performance were significant and negative for all memory tests, although considerably stronger in the imagery group; cued recall accuracy (imagery: $r(111) = -.50, p < .001, BF_{10} > 1000$, control: $r(112) = -.19, p = .042, BF_{10} = 0.58$), associative recognition d' (imagery: $r(54) = -.68, p < .001, BF_{10} > 1000$, control: $r(56) = -.25, p = .06, BF_{10} = 0.59$), and order recognition d' (imagery: $r(55) = -.43, p = .001, BF_{10} = 23.45$, control: $r(54) = -.42, p = .0015, BF_{10} = 16.04$). Thus, participants with high initial performance may have already found a strategy as effective as interactive imagery, explaining the weaker effectiveness of our manipulation.

The relationship between order recognition and cued recall: within-subject analyses

Kato and Caplan (2017) tested each word pair with cued recall, and then either associative or order recognition depending on condition. In their study, order recognition performance for correctly recalled pairs was significantly better than for incorrectly recalled pairs, but well below this same difference for associative recognition. The following analyses test if instructed imagery instructions in experiment 1 modified these patterns.

As a reminder, for performance on order and associative recognition tests, we measured d' = z(hit rate) $-$ z(false alarm rate). Whenever hit or false alarm rate were zero or one, one-half an observation was added or subtracted to avoid infinities. Because of the correction d'_{max}, or the maximum possible d' value, depends on the number of trials included. We computed d'_{max} based on a (corrected) a hit rate of one, and a false alarm rate of zero, as a reference for the order and associative recognition analyses separated by correctness in cued recall. Because participants varied in the amount of correct and incorrect cued recall trials, d'_{max} also varied across participants. These d'_{max} values, alongside recognition performance separately computed for correctly versus incorrectly recalled pairs, are plotted in Figure 2.18.

To test if order recognition had less dependence on cued recall correctness than associative recognition, we subtracted performance for incorrectly recalled pairs from performance for correctly recalled pairs, for both order recognition and associative recognition,[12] to obtain difference scores for each task for both groups and in both instruction phases. A mixed, repeated-measures ANOVA was performed on this difference score measure, with the design Group (imagery, control) × Instruction phase (pre-instruction, post-instruction) × Task (associative recognition, order recognition). This analysis returned a significant main effect of Task, $F(1,188) = 9.36$, $MSE = 9.91$, $p = .003$, $\eta_p^2 = 0.047$, $BF_{\text{inclusion}} = 3.47$. All other effects were non-significant (all $p > .07$, all $BF_{\text{inclusion}} < 0.3$), indicating that associative recognition had significantly larger difference scores than order recognition, regardless of group or instruction phase. Thus, our results

[12] In associative recognition, recombined probes contain items that were not paired at study. In our study we identified correctly recalled pairs using the recall outcome of the left item in the probe. It would also be possible to base this measure on recall outcome of the right item, but Kato & Caplan (2017) found that this made little difference.

replicate the weaker coupling of order recognition to cued recall found in Kato & Caplan (2017).

However, visual inspection of Figure 2.18 shows that due to differences in trial counts, and the correction to avoid infinities (see Methods), the maximum possible d' value was not constant across conditions. As a second way to ask about the relative coupling of order recognition to cued recall accuracy, we next took $d'_{\max}$ into account. To test if associative recognition was closer to $d'_{\max}$ as compared to order recognition, we subtracted each participant's observed d' from their $d'_{\max}$, for both associative and order recognition for correctly and incorrectly recalled pairs, and for both groups and instruction phases. Independent samples t-tests indicated that associative recognition was closer to $d'_{\max}$ than order recognition for correctly recalled pairs, in both groups and both instruction phases (all $p < .001$, $BF_{10} > 1000$). For incorrectly recalled pairs, this same difference between observed d' and $d'_{\max}$ was not significant in the control and imagery group pre-instruction (both $p > .34$, $BF_{10} < 0.31$), and the control group post-instruction, $t(109) = -1.43$, $p = .16$, $BF_{10} = .50$, but in the imagery group post-instruction, associative recognition for incorrectly recalled pairs was significantly closer to $d'_{\max}$ than order recognition, $t(104) = -2.89, p = .005$, $BF_{10} = 7.92$. In sum, when taking the maximum measurable d' into account, the relationship between order recognition and cued-recall is well below perfect.

To examine if the coupling between order recognition and cued recall was zero, as would be expected for order-absent models, we ran paired-samples t-tests between order recognition for correctly recalled pairs, and order recognition for incorrectly recalled pairs. Order recognition was significantly higher for correctly recalled pairs for both groups, and in both instruction phases (all $p < .001$, $BF_{10} > 32$), indicating non-zero coupling between order recognition and cued recall.

Imagery instructions increased associative recognition performance overall. Paired t-tests indicated that associative recognition was significantly higher for both correctly recalled pairs, $t(49) = 4.92, p < .001$, $BF_{10} > 1000$ and incorrectly recalled pairs, $t(50) = 4.03, p < .001$, $BF_{10} = 125.51$. Order recognition performance for correctly recalled pairs and incorrectly recalled pairs did not significantly change after the imagery instruction (both $p > .23$, $BF_{10} < 0.3$).

In sum, just as in Kato and Caplan (2017), order recognition d' had a significant dependence on cued recall correctness; however, this relationship was significantly smaller than observed between associative recognition and cued recall, and order recognition performance was significantly below maximum, even for correctly recalled pairs. Order recognition did not have maximal relationship with cued recall (as perfect-order models would predict), nor a null relationship with cued recall (as order-absent models would predict), but a mid-range relationship inconsistent with all model accounts. Imagery instructions did not affect these patterns.

2.7.2 Experiment 2

Correlations between visual imagery measures and memory performance

Tables 2.8– 2.10 report each correlation between visual imagery measures (PFT and VVIQ) and performance in cued recall, associative recognition and order recognition tasks.

Self-report on strategy use

At the end of the session in experiment 2, participants answered three strategy-use questions on a scale of one (never) to five (always), in succession; Q1) "When studying the word pairs, how often did you imagine an image (in your mind's eye)?", Q2) "When studying the word pairs, how often did you imagine the word pairs interacting with each other?", Q3) "When studying the word pairs, how often did you incorporate order into your mental image? ".

To check if the Mental Imagery Frequency rating had a relationship to the effect of interactive imagery, we conducted A two-way ANOVA on post-minus-pre cued recall accuracy with the design Group [3] × Mental Imagery Frequency rating [5]. This returned a significant effect of Group, $F(2,410) = 11.17, MSE = 0.51$, $p < .001$, $\eta_p^2 = 0.052$, $BF_{\text{inclusion}} > 1000$. There was a significant effect of Mental Imagery Frequency rating, $F(4,410) = 2.39$, $MSE = 0.11$, $p = .05$, $\eta_p^2 = 0.023$, $BF_{\text{inclusion}} = 0.14$, although a Bayesian ANOVA returned supported null evidence. Nonetheless, we cautiously followed up with post-hoc tests, which indicated post-minus-pre cued recall accuracy for rating 4 was nearly significantly larger than rating one (never), $p_{\text{tukey}} = .065$, providing some evidence for a imagery strategy benefit (collapsed across groups), although all other post-hoc tests were not significant ($p_{\text{tukey}} > .14$).

To check if the Interactivity Frequency rating had a relationship to the effect of interactive imagery, we conducted a two-way ANOVA on post-minus-pre cued recall accuracy with the design Group [3] × Interactivity Frequency rating [5]. This returned a significant effect of Group, $F(2,410) = 9.25$, $MSE = 0.43$, $p < .001$, $\eta_p^2 = 0.043$, $BF_{\text{inclusion}} > 1000$, but a non-significant, supported null effect for Group and interaction with the effect of Interactivity Frequency rating (both $p > .24$, $BF_{\text{inclusion}} < 0.3$), suggesting that self-reported imagining of interactivity between words did not affect cued recall accuracy, matching results from the subjectively scored free form responses in experiment 1.

We also checked if self-reported frequency of incorporating order into the mental image (rated never to always) had an effect on order recognition d'. We subtracted pre-instruction from post-instruction order recognition d' and performed a two-way ANOVA on this measure, with the design Group[3] × Order Incorporation rating[5]. All effects and interactions were not significant and supported null (all $p > .18$, $BF_{\text{inclusion}} < 0.3$), indicating that self-reported incorporation of within-pair order at study did not affect order recognition performance.

Mid-session strategy instruction comprehension question

Immediately after participants received a strategy instruction in experiment 2, they were asked to describe what they had just been asked to do. To quantify the degree to which participants understood instructions, these responses were rated by two separate coders blinded to group (KA and JT); First, based on the experimental group the coder thought the participant belonged to, 0) I don't know/Empty,[13] 1) standard-imagery, 2) top-bottom imagery, 3) actor-object imagery. These ratings were then compared to the actual group of the participant and scored either correct or incorrect, returning what we term "Group Identification rating" in following analyses. Second, responses were scored on whether participants understood the instruction, which we term "Instruction Comprehension rating", 0) Zero understanding/Empty, 1) Somewhat understands, 2) Understands. All 433 participants were included. After initial coding, inter-rater reliability was substantial for Group Identification ratings (Cohen's $\kappa = 0.81$), but low for Instruction Comprehension ratings (Cohen's

[13] Empty indicating no response entered.

$\kappa = 0.58$). Thus, raters met and came to consensus for all disagreeing ratings, and these are the values we report. To check if the ineffectiveness of top-bottom imagery instructions was due to lack of comprehension, we repeated analyses from the main text that were performed on mean cued recall, associative recognition, and order recognition, on a subset of participants with the highest Instruction Comprehension rating i.e., "Understands" , and separately, on participants with correct Group Identification ratings.

Cued recall accuracy Restricted to participants with the highest Instruction Comprehension rating (e.g., "Understands"), a mixed ANOVA was performed on cued recall accuracy with design Group × Instruction phase. Following analysis of all participants regardless of rating (reported in the main text), there was a significant main effect of Instruction phase, $F(1,248) = 51.16$, $MSE = 1.26$, $p < .001$, $\eta_p^2 = 0.17$, $BF_{\text{inclusion}} > 1000$, and significant Group × Instruction phase interaction, $F(2,248) = 19.29$, $MSE = 0.48$, $p < .001$, $\eta_p^2 = 0.14$, $BF_{\text{inclusion}} > 1000$. Simple effects indicated a significant increase in performance post-instruction in both the actor-object, and standard-imagery group (both $p < .001$, $BF_{10} > 1000$), but a supported null difference in the top-bottom imagery group ($p = .52$, $BF_{10} < 0.3$). Additionally, there was a supported null difference between Group pre-instruction ($p = .19$, $BF_{10} < 0.3$), but significant post-instruction ($p < .001$, $BF_{10} = 77.33$). Follow up t-tests on the significant post-instruction Group difference indicated a non-significant, supported null difference between the standard and actor-object imagery, $p = .34$, $BF_{10} < 0.3$, and that top-bottom imagery was significantly worse than standard-imagery ($p < .001$, $BF_{10} = 142.12$) and actor-object imagery ($p = .007$, $BF_{10} = 5.31$). Next, restricted to participants with correct Group Identification ratings, a mixed ANOVA was performed on cued recall accuracy with design Group × Instruction phase. Again, there was a significant main effect of Instruction phase, $F(1,298) = 46.09$, $MSE = 1.15$, $p < .001$, $\eta_p^2 = 0.13$, $BF_{\text{inclusion}} > 1000$, and significant Group × Instruction phase interaction, $F(2,298) = 25.87$, $MSE = 0.65$, $p < .001$, $\eta_p^2 = 0.15$, $BF_{\text{inclusion}} > 1000$. Simple effects indicated a significant increase in performance post-instruction in both the actor-object, and standard-imagery group (both $p < .001$, $BF_{10} > 375$), but a supported null difference in the top-bottom imagery group ($p = .17$, $BF_{10} < 0.3$). Additionally, there was a supported null difference between Group pre-instruction ($p = .12$, $BF_{10} < 0.3$), but significant post-instruction ($p < .001$, $BF_{10} = 2170.09$). Follow up t-tests on the significant post-instruction Group difference indicated a non-significant, nearly supported null difference between the standard and actor-object imagery, $p = .18$, $BF_{10} = 0.38$, and that top-bottom imagery was significantly worse than standard-imagery ($p < .001$, $BF_{10} > 1000$) and actor-object imagery ($p = .004$, $BF_{10} = 8.78$). In sum, even when restricted to participants who demonstrated high instruction comprehension, top-bottom instructions were ineffective to improve cued recall accuracy, while actor-object and standard-imagery instructions improved performance to a similar degree.

Associative recognition Restricted to participants with the highest Instruction Comprehension rating (e.g., "Understands"), a mixed ANOVA on associative recognition d', with design Group × Instruction phase returned significant main effects of Instruction phase, $F(1,105) = 27.86$, $MSE = 13.31$, $p < .001$, $\eta_p^2 = 0.21$, $BF_{\text{inclusion}} > 1000$, and significant Group × Instruction phase interaction, $F(2,105) = 8.49$, $MSE = 5.58$, $p < .001$, $\eta_p^2 = 0.14$, $BF_{\text{inclusion}} = 43.18$. Simple effects indicated that associative recognition performance increased post-instruction in both the actor-object group ($p < .001$, $BF_{10} = 515.28$) and standard-imagery group ($p < .001$, $BF_{10} = 654.99$) groups, while the top-bottom group had a supported null difference between instruction phases ($p = .92$, $BF_{10} < 0.3$). Simple effects with the factor Group returned a supported null difference pre-instruction ($p = .37$, $BF_{\text{inclusion}} < 0.3$), but a significant difference post-instruction ($p = .004$, $BF_{\text{inclusion}} = 9.92$). Follow-up t-tests on the post-instruction group difference indicate that actor-object and standard-imagery had a supported null difference ($p = .72$, $BF_{10} < 0.3$), but both groups performed significantly better than the top-bottom group ($p = .004$, $BF_{10} = 9.92$ and $p = .01$, $BF_{10} = 4.18$ respectively). Restricted to participants with correct Group Identification ratings, a mixed ANOVA on associative recognition d', with design Group × Instruction phase returned significant main effects of Instruction phase, $F(1,128) = 20.18$, $MSE = 13.61$, $p < .001$, $\eta_p^2 = 0.14$, $BF_{\text{inclusion}} > 1000$, and significant Group × Instruction phase interaction,

$F(2,128) = 9.62$, $MSE = 6.49$, $p < .001$, $\eta_p^2 = 0.13$, $BF_{\text{inclusion}} = 183.72$. Simple effects indicated that associative recognition performance increased post-instruction in both the actor-object group ($p = .002$, $BF_{10} = 18.24$) and standard-imagery group ($p < .001$, $BF_{10} > 1000$) groups, while the top-bottom group had a supported null difference between instruction phases ($p = .47$, $BF_{10} < 0.3$). Simple effects with the factor Group returned a supported null difference pre-instruction ($p = .44$, $BF_{\text{inclusion}} < 0.3$), but a significant difference post-instruction ($p < .001$, $BF_{\text{inclusion}} = 54.37$). Follow-up t-tests on the post-instruction Group difference indicate that actor-object and standard-imagery had a supported null difference ($p = .90$, $BF_{10} < 0.3$), but both groups performed significantly better than the top-bottom group ($p = .002$, $BF_{10} = 14.67$ and $p = .001$, $BF_{10} = 27.53$ respectively). In sum, even in participants selected for high Instruction Comprehension ratings top-bottom instructions were significantly less effective for associative recognition compared to standard and actor-object imagery instructions.

Order recognition Restricted to participants with the highest Instruction Comprehension rating (e.g., "Understands"), a mixed ANOVA on order recognition d', with design Group × Instruction phase returned significant main effects of Instruction phase, $F(1,140) = 12.98$, $MSE = 5.98$, $p < .001$, $\eta_p^2 = 0.09$, $BF_{\text{inclusion}} = 22.87$, Group, $F(2,140) = 3.55$, $MSE = 4.21$, $p = .03$, $\eta_p^2 = 0.05$, $BF_{\text{inclusion}} = 1.33$ (although Bayesian analyses returned inconclusive evidence for Group), and a non-significant effect of Group × Instruction phase ($p = .18$, $BF_{\text{inclusion}} = 0.71$). Following up on the main effect of Group with post-hoc tests returns a significant difference between the standard and top-bottom group ($p_{\text{tukey}} = .026$), and non-significant differences between the actor-object and the other two groups (both $p_{\text{tukey}} > .19$); however, because these Group differences are only significant when collapsing across pre- and post-instruction phases, and the interaction Group × Instruction phase was not significant, our results still suggest that order emphasizing strategy instructions did not have an advantage over standard-imagery instructions for order recognition, even when restricting to participants with the highest Instruction Comprehension rating. Restricted to participants with correct Group Identification ratings, a mixed ANOVA on order recognition

d', with design Group × Instruction phase returned significant main effects of Instruction phase, $F(1,167) = 8.96$, $MSE = 4.62$, $p = .003$, $\eta_p^2 = 0.05$, $BF_{\text{inclusion}} = 4.57$, but the main effect and interaction involving Group were not significant and supported null (both $p > .19$, $BF_{\text{inclusion}} < 0.3$).

In sum, even when accounting for Instruction Comprehension, order-emphasizing instructions did not improve order recognition more than standard-imagery instructions.

Aphantasia case studies

Out of 433 total participants, 120 participants self-identified as aphantasic with the end-of-session aphantasia identification question. However, because participants in experiment 2 did not complete a previous mass-testing questionnaire, we could not verify consistency across multiple responses. Thus, we moved directly to the in-session VVIQ criteria stated in experiment 3. Among the 120 yes responders to the aphantasia question, four participants met our moderate aphantasia criteria of 73/80, and one participant met our extreme criteria of 80/80. These five participants are reported as case studies in Table 2.11.

Among these five participants, participant 3 received standard interactive imagery instructions and exhibited a 68% increase in cued recall accuracy post-instruction, consistent with results from experiment 3 that interactive imagery instructions were just as effective for aphantasics. Three out of five participants (1, 4, 5), including one extreme aphantasic, received top-bottom imagery instructions, and all exhibited essentially no change to cued recall accuracy (+3.1%), or a substantial reduction, consistent with findings in the larger sample that top-bottom instructions were ineffective for cued recall. Participant 2 received actor-object instructions and exhibited a substantial reduction in cued recall performance, but a large increase in order recognition, a pattern that should be followed up in a larger sample of aphantasics.

Scatter plots of log-odds cued recall versus order and associative recognition

Figures 2.19 and 2.20 are scatter-plots of log-odds transformed cued recall accuracy versus both order and associative recognition d'.

The relationship between order recognition and cued recall: within-subject analyses

As we report below, within-subject OR-CR versus AR-CR analyses diverged somewhat from results in experiment 1. This may have been because associative recognition performance separated by correct versus incorrectly recalled pairs was especially sensitive to low trial counts for recombined trials (see experiment 2 methods). Thus, the following analyses involving associative recognition should be interpreted with some caution. Additionally, in the pre-registration for experiment 2, an analysis of recognition d'_{max} for correctly and incorrectly recalled pairs was planned; However, instead of d'_{max}, we analyzed hit rates and false alarm rates.

To quantify the *within-subject* relationship between order recognition and cued recall, we subtracted each participant's recognition (order and associative) performance for incorrectly recalled pairs from performance for correctly recalled pairs, and performed analyses on this difference. A mixed, repeated-measures ANOVA was performed on this d' difference measure with the design Group (standard, actor-object, top-bottom) × Instruction phase (pre-instruction, post-instruction) × Task (associative recognition, order recognition). All effects and interactions were not significant and supported null (all $p > .41$, $BF_{\text{inclusion}} < 0.3$); however, the effect of Task nearly reached significance $F(1,374) = 4.32$, $MSE = 5.15$, $p = .038$, $\eta_p^2 = 0.011$, $BF_{\text{inclusion}} = 0.35$, although with supported null evidence in the Bayesian analysis. Thus, the expected effect of Task (which indicates a difference between associative and order recognition's relationship to cued recall), was not observed. However, the near significance of Task ($p = .038$) suggests the conclusion of a supported null effect in the Bayesian analysis must be interpreted with some caution.[14] The nearly significant effect of Task led us to break analyses down into hit rates and false alarm rates, to check if the expected patterns would be observed at these levels.

Hit rates. A mixed, repeated-measures ANOVA was performed on the *hit rate* difference measure with the design Group (standard, actor-object, top-bottom) × Instruction phase (pre-instruction, post-instruction) × Task (associative recognition, order recognition), and a significant main effect

[14] Additionally, when applying the d' correction suggested by Hautus (1995), the basic effect of Task (indicating a smaller OR-CR relationship, compared to the AR-CR relationship) replicated.

of Task $F(1,402) = 8.82$, $MSE = 0.43$, $p = .003$, $\eta_p^2 = 0.021$, $BF_{\text{inclusion}} = 2.79$, indicating that associative recognition had significantly larger difference in hit rate than order recognition. The main effect of Instruction phase was also significant, $F(1,402) = 6.36$, $MSE = 0.24$, $p = .01$, $\eta_p^2 = 0.016$, $BF_{\text{inclusion}} = 0.88$, indicating that hit rate difference reduced post-instruction overall, although the Bayesian analysis indicated weak evidence for this effect. All other main effects and interactions, and most importantly those involving Group, were not significant and supported null (all $p > .12$, $BF_{\text{inclusion}} < 0.3$), suggesting that there was no effect of either of the three imagery instructions on the relationship between order recognition and cued recall. In sum, analyses of hit rates were consistent with the weaker coupling of order recognition to cued recall found in Kato & Caplan (2017). Paired-samples t-tests indicated order recognition hit rate were significantly higher for correctly recalled pairs, compared to incorrectly recalled pairs for all groups, and in both instruction phases (all $p < .006$, $BF_{10} > 5.42$), indicating non-zero coupling between order recognition and cued-recall contrary to order-absent models.

False alarm rates. A mixed, repeated-measures ANOVA was performed on the *false alarm rate* difference measure with same design: Group (standard, actor-object, top-bottom) × Instruction phase (pre-instruction, post-instruction) × Task (associative recognition, order recognition). There significant main effect of Instruction phase $F(1,390) = 11.71$, $MSE = 0.91$, $p < .001$, $\eta_p^2 = 0.029$, $BF_{\text{inclusion}} = 6.21$, indicating an overall reduction in the difference between false alarm rates for correct and incorrectly recalled pairs post-instruction. All other effects were not significant, and supported null ($p > .07$, $BF_{\text{inclusion}} < 0.3$). Thus, the dependence of order recognition false alarm rates on cued recall was not significantly different than the dependence of associative recognition false alarm rates. Paired-samples t-tests indicated order recognition false alarm rates were lower for correctly recalled pairs compared to incorrectly recalled pairs for all groups, and in both instruction phases (all $p < .02$, $BF_{10} > 1.63$).

In sum, despite divergence at the level of false alarm rates and d' (which may have been especially affected by low trial counts for recombined trials), patterns in hit rates still challenge both perfect-order and order-absent mathematical models. Perfect-order models cannot account for the lesser dependence of order recognition hit rates on cued recall correctness, compared to associative recognition. Order-absent models cannot account for the significant effect of cued recall correctness on order recognition hit rates, and false alarms. Importantly, there was no evidence that any instruction had an effect on these patterns.

2.7.3 Experiment 3

Correlations between visual imagery measures and memory performance

Table 2.13 reports each correlation between visual imagery measures (PFT and VVIQ) and performance in the cued recall task.

Self-report on strategy use

At the end of the session in experiment 3, participants were asked to "describe how you studied the word pairs, whether or not that included the use of visual imagery as instructed, in a short one or two sentence response." These responses were rated by two coders, blinded to condition, for two measures of interest. Firstly, rated either; 1) response includes imagery, 2) response explicitly excludes imagery, 3) response leaves open the possibility of imagery but was not explicit. Next, each response was rated for whether it referred to interactivity or connection between words (yes/no). Of the 122 participants, 13 provided no response, and were omitted from this analysis. After initial coding, inter-rater reliability was substantial for Imagery Reference scoring (Cohen's $\kappa = 0.76$), but somewhat lower for Interactivity Reference scoring (Cohen's $\kappa = 0.41$). As a result, we encouraged the coders to meet and come to consensus on disagreeing responses. Coders were able to come to a consensus for all responses, and these are the ratings we report. One participant completed the experiment after the coding was completed and was coded based on the same coders' consensus. First, there was a trend towards aphantasics referring to imagery (54% of responses) less than inconsistent responders (74% of responses) and less than non-aphantasics (78% of responses) but this was not significant, $\chi^2(4, N = 110) = 6.75$, $p = .15$.

Next, to test if change in cued recall accuracy (from pre-instruction to post-instruction) was affected by imagery-report ratings, we ran an ANOVA on change in cued recall accuracy, with design Group[3] × Imagery rating[3]. There was a significant main effect of Imagery rating, $F(2,107) = 3.78, MSE = 0.13, p = .026, n_p^2 = 0.07, BF_{\text{inclusion}} = 2.78$, but the effects of Group and the interaction were not significant (both $p > .77$, $BF_{\text{inclusion}} < 0.3$). Post-hoc Tukey tests indicated that participants who referred to imagery exhibited a significantly higher change in cued recall accuracy than participants who explicitly excluded imagery (rating 2), $p_{\text{tukey}} = .025$, but were not significantly different than participants who left open the possibility of imagery but were not explicit (rating 3), $p_{\text{tukey}} = .46$. Additionally, participants with a rating of two were not significantly different than participants with a rating of three, $p_{\text{tukey}} = .56$. A smaller proportion of aphantasics referred to imagery in their self-report, suggesting that they would exhibit lower memory performance; however, the findings above favoured null differences between self-identified aphantasics and non-aphantasics in cued recall performance. Thus, the imagery self-report effect was evidently not large enough to cause meaningful differences in aphantasic memory performance.

Consistent aphantasics also referred to interactivity (54%), less than inconsistent responders (76% of responses), and consistent non-aphantasics (75% of responses), but this was also not significant, $\chi^2(2, N = 110) = 4.18$, $p = .12$. An ANOVA on Group[3] × Interactivity rating[2] returned all non-significant, supported null effects (all $p > .09$, $BF_{\text{inclusion}} < 0.3$). In sum, although there is a trend towards aphantasics referring to interactivity less than other groups, this rating had little relationship to objective effectiveness of interactive imagery instructions.

Gender and interactive imagery effects

We could find no analysis of the influence of gender on interactive imagery effects in previous literature. This motivated us to test whether self-reported gender could influence the general patterns observed in this study. For participants from experiment 3, we gathered gender-identification responses from the Winter 2021 mass questionnaire (see experiment 3 methods). Note, eight participants recruited to experiment 3 did not fill out the Winter 2021 mass questionnaire, because they were recruited to the study through their Fall 2020 questionnaire responses. Thus, we did not include their data in the following analyses. Additionally, one participant self-identified as non-binary, and one participant did not wish to disclose. Because there was only one participant for each of these groups, we could not include these participants in the following statistical analyses.

A mixed ANOVA on cued recall accuracy with the design Instruction phase (pre-instruction, post-instruction) × Self-reported gender (male, female), a supported null effect of Self-reported gender and the interaction Instruction phase × Self-reported gender (both $p > .41$, $BF_{\text{inclusion}} < 0.3$). Thus, we found no evidence for that gender influenced the effectiveness of imagery instructions.

Additionally, independent samples t-tests between self-identified males and females returned non-significant supported null differences in VVIQ ratings, PFT accuracy and PFT response times (all $p > .49$, $BF_{10} < 0.3$), indicating mean values in our visual imagery measures did not differ based on gender.

Furthermore, correlations between the VVIQ, PFT accuracy, PFT response times to the post-minus-pre cued recall accuracy were not significant, and either weak or supported null for self-reported females (VVIQ: $r(80) = -.005, p = .96, BF_{10} = 0.14$, PFT accuracy: $r(80) = -.17, p = .12, BF_{10} = 0.46$, PFT response times: $r(80) = -.17, p = .14, BF_{10} = 0.41$), and for self-reported males, (VVIQ: $r(28) = -.035, p = .86, BF_{10} = 0.23$, PFT accuracy: $r(28) = .002, p = .99, BF_{10} = 0.23$, PFT response times: $r(28) = .10, p = .61, BF_{10} = 0.26$). Thus, regardless of gender, there was no relationship between interactive imagery effectiveness and individual differences in visual imagery.

Mass questionnaire VVIQ and test-retest reliability The VVIQ was included in both Fall 2020 and Winter 2021 mass questionnaires. Test-retest reliability between the Winter 2021 mass questionnaire administration, and our in-session administration was good, $r(110) = .88$. Reliability between the Fall 2020 administration to in-session ratings, $r(78) = .60$, and Winter 2021 administration, $r(740) = .59$, was somewhat lower, which may warrant caution in interpreting Fall 2020 VVIQ ratings. However, all analyses in this study are based on the in-session VVIQ administration, and the good reliability between Winter 2021 and in-session ratings suggest that our in-session VVIQ ratings were reliable.

Scatter plots of log-odds cued recall versus order and associative recognition

Figure 2.15 depicts scatter plots of log-odds transformed cued recall accuracy versus both order and associative recognition d'.

2.7.4 All Experiments

As an alternative way to test the relationship between imagery vividness/ability and the effectiveness of cued recall, we computed correlations between post-minus-pre-instruction memory performance, to VVIQ ratings (all experiments) and PFT accuracy/response times (experiments 1 and 3). These are reported in Tables 2.16–2.18. In general, these correlations were either weak or supported null, supporting the conclusions in the main manuscript that individual differences in visual imagery ability do not relate to the effectiveness of interactive imagery. In Experiment 1, increased PFT response time predicted a greater change in cued recall accuracy, and a greater change in associative recognition in the imagery group (Table 2.16), although we explain in the main text how longer PFT response times more likely indicate increased effort/engagement rather than imagery skill. Also in experiment 1, there was significant correlation between VVIQ ratings and the change in associative recognition in the imagery group (Table 2.17); however, this correlation was not replicated in experiment 2.

Table 2.5: Experiment 1: Correlations between cued recall accuracy and visual imagery measures.

	VVIQ ratings			PFT accuracy			PFT response time		
	r	p	BF_{10}	r	p	BF_{10}	r	p	BF_{10}
Pre-instruction: Imagery	.02	.86	0.12	.20	.03*	1.21	.01	.93	0.12
Pre-instruction: Control	−.07	.44	0.16	.27	.004*	6.61	.17	.08	0.53
Pre-instruction Fisher test (Imagery versus Control)	$z = 0.67$, $p = .50$			$z = 0.48$, $p = .63$			$z = 1.17$, $p = .24$		
Post-instruction: Imagery	−.15	.10	0.44	.28	.002*	10.90	.27	.004*	7.49
Post-instruction: Control	.02	.87	0.12	.24	.01*	2.86	.12	.22	0.25
Post-instruction Fisher test (Imagery versus Control)	$z = 1.27$, $p = .20$			$z = 0.36$, $p = .72$			$z = 1.20$, $p = .23$		

indicates significance at .05.

Table 2.6: Experiment 1: Correlations between associative recognition d' and visual imagery measures.

	VVIQ ratings			PFT accuracy			PFT response time		
	r	p	BF_{10}	r	p	BF_{10}	r	p	BF_{10}
Pre-instruction: Imagery	.03	.80	0.17	.18	.18	0.40	.15	.28	0.30
Pre-instruction: Control	−.12	.37	0.24	.35	.008*	5.33	.18	.19	0.38
Pre-instruction Fisher test (Imagery versus Control)	$z = 0.80$, $p = .42$			$z = 0.94$, $p = .35$			$z = 0.15$, $p = .88$		
Post-instruction: Imagery	−.44	< .001*	44.10	.34	.011*	3.76	.32	.017*	2.74
Post-instruction: Control	−.04	.78	0.17	.38	.003*	11.12	.18	.17	0.41
Post-instruction Fisher test (Imagery versus Control)	$z = 2.25$, $p = .024$*			$z = 0.26$, $p = .79$			$z = 0.76$, $p = .45$		

indicates significance at .05.

Table 2.7: Experiment 1: Correlations between order recognition d' and visual imagery measures.

	VVIQ ratings			PFT accuracy			PFT response time		
	r	p	BF_{10}	r	p	BF_{10}	r	p	BF_{10}
Pre-instruction: Imagery	.11	.43	0.22	.21	.12	0.55	$-.04$	.76	0.17
Pre-instruction: Control	$-.05$	.71	0.18	.43	.001*	30.98	.30	.027*	1.82
Pre-instruction Fisher test (Imagery versus Control)	$z = 0.82$, $p = .41$			$z = 1.25$, $p = .21$			$z = 1.79$, $p = .07$		
Post-instruction: Imagery	$-.03$	.84	0.17	.16	.24	0.33	.22	.10	0.60
Post-instruction: Control	.19	.16	0.45	.41	.002*	18.43	.37	.005*	8.27
Post-instruction Fisher test (Imagery versus Control)	$z = 1.15$, $p = .25$			$z = 1.40$, $p = .16$			$z = 0.88$, $p = .38$		

indicates significance at .05.

Table 2.8: Experiment 2: Correlations between cued recall accuracy and VVIQ ratings.

	VVIQ ratings		
	r	p	BF_{10}
Pre-instruction: Standard-Imagery	$-.07$	.40	0.14
Pre-instruction: Actor-Object Imagery	$-.07$	.43	0.15
Pre-instruction: Top-Bottom Imagery	$-.18$	.03*	1.09
Post-instruction: Standard-Imagery	$-.05$	.52	0.12
Post-instruction: Actor-Object Imagery	$-.13$	.16	0.14
Post-instruction: Top-Bottom Imagery	$-.12$	.17	0.27

indicates significance at .05.

Table 2.9: Experiment 2: Correlations between associative recognition d' and VVIQ ratings.

	VVIQ ratings		
	r	p	BF_{10}
Pre-instruction: Standard-Imagery	.01	.91	0.15
Pre-instruction: Actor-Object Imagery	−.02	.88	0.17
Pre-instruction: Top-Bottom Imagery	−.11	.36	0.26
Post-instruction: Standard-Imagery	−.06	.59	0.29
Post-instruction: Actor-Object Imagery	.04	.79	0.26
Post-instruction: Top-Bottom Imagery	.02	.86	0.15

indicates significance at .05.

Table 2.10: Experiment 2: Correlations between order recognition d' and VVIQ ratings.

	VVIQ ratings		
	r	p	BF_{10}
Pre-instruction: Standard-Imagery	−.03	.80	0.14
Pre-instruction: Actor-Object Imagery	−.002	.99	0.14
Pre-instruction: Top-Bottom Imagery	−.07	.59	0.17
Post-instruction: Standard-Imagery	.03	.80	0.14
Post-instruction: Actor-Object Imagery	.14	.22	0.30
Post-instruction: Top-Bottom Imagery	−.05	.67	0.17

indicates significance at .05.

Table 2.11: Experiment 2: Group, condition, change in cued recall accuracy, change in recognition d', and VVIQ ratings for yes responders to the end-of-session aphantasia question who scored higher than 73 on the VVIQ.

Participant	Group	Changed in cued recall accuracy	Condition	Change in Recognition d'	VVIQ (out of 80)
1	Top-Bottom	-12.5%	Order recognition	-0.17	80
2	Actor-Object	-25%	Order recognition	$+1.00$	79
3	Standard	$+68\%$	Order recognition	$+0.93$	77
4	Top-Bottom	$+3.1\%$	Associative recognition	$+0.07$	77
5	Top-Bottom	-59%	Associative recognition	-0.61	74

Table 2.12: Experiment 1: Correlations between log-odds cued recall accuracy and both associative and order recognition, broken down by direction of cued recall test for recognition probes.

	Forward cued recall test		Backward cued recall test	
	r	p	r	p
Pre-instruction: Imagery Associative Recognition	.80	$< .001$	.68	$< .001$
Pre-instruction: Imagery Order recognition	.38	.004	.33	.012
Fisher test (OR versus AR)	$z = 3.71$, $p < .001$		$z = 2.52$, $p = .012$	
Pre-instruction: Control Associative Recognition	.67	$< .001$	.77	$< .001$
Pre-instruction: Control Order recognition	.47	$< .001$	.24	.08
Fisher test (OR versus AR)	$z = 1.62$, $p = .10$		$z = 3.98$, $p < .001$	
Post-instruction: Imagery Associative Recognition	.71	$< .001$	.53	$< .001$
Post-instruction: Imagery Order recognition	.30	.021	.17	.21
Fisher test (OR versus AR)	$z = 2.92$, $p = .0035$		$z = 2.19$, $p = .029$	
Post-instruction: Control Associative Recognition	.80	$< .001$	.72	$< .001$
Post-instruction: Control Order recognition	.41	.0019	.23	.085
Fisher test (OR versus AR)	$z = 3.50$, $p < .001$		$z = 3.52$, $p < .001$	

Table 2.13: Experiment 3: Correlations between cued recall accuracy to VVIQ ratings, PFT accuracy and PFT response time.

	VVIQ ratings			PFT accuracy			PFT response time		
	r	p	BF_{10}	r	p	BF_{10}	r	p	BF_{10}
Pre-instruction: Total participants	−.06	.48	0.14	.46	< .001*	> 1000	.23	.01*	2.75
Pre-instruction: Consistent aphantasics	.09	.67	0.27	.55	.005*	10.67	.33	.10	0.88
Pre-instruction: Consistent non-aphantasics	.18	.30	0.36	.36	.04*	1.75	−.03	.86	0.22
Pre-instruction: Inconsistent responders	−.27	.03*	1.42	.49	< .001*	481.04	.30	.02*	2.78
Post-instruction: Total participants	−.08	.40	0.16	.39	< .001*	> 1000	.21	.02*	1.45
Post-instruction: Consistent aphantasics	−.14	.50	0.31	.54	.005*	9.77	.51	.009*	6.39
Post-instruction: Consistent non-aphantasics	−.15	.40	0.30	.55	< .001*	49.61	.01	.96	0.21
Post-instruction: Inconsistent responders	−.05	.68	0.17	.28	.03*	1.76	.21	.10	0.59

indicates significance at .05.

Table 2.14: Experiment 3: Number of participants whose free form strategy response referred to imagery, did not refer to imagery, or left open the possibility of imagery, as rated by coders blinded to group. Note that certain participants did not include a free form response, accounting for fewer participants in this table than the total sample size.

Response rating	Inconsistent responders	Consistent aphantasics	Consistent non-aphantasics
Includes imagery	40	13	25
Explicitly excludes imagery	5	7	4
Leaves open the possibility of imagery	9	4	3

Table 2.15: Experiment 3: number of participants whose free form strategy response referred to interactivity or did not refer to interactivity, rated by coders blinded to group. Note that certain participants did not include a free form response, accounting for fewer participants in this table than the total sample size.

Response rating	Inconsistent responders	Consistent aphantasics	Consistent non-aphantasics
Does not refers to interactivity	13	11	8
Refer to interactivity	41	13	24

Table 2.16: All experiments: correlations between post-minus-pre instruction cued recall accuracy and visual imagery measures.

	VVIQ ratings			PFT accuracy			PFT response time		
	r	p	BF_{10}	r	p	BF_{10}	r	p	BF_{10}
Experiment 1: Imagery group	−.15	.11	0.41	.09	.36	0.18	.23	.01*	2.47
Experiment 1: Control group	.11	.25	0.22	.02	.82	0.12	−.03	.74	0.12
Experiment 2: Top-bottom imagery	.07	.44	0.14		N/A			N/A	
Experiment 2: Actor-object imagery	−.06	.50	0.14		N/A			N/A	
Experiment 2: Standard interactive imagery	.02	.81	0.10		N/A			N/A	
Experiment 3: Consistent aphantasics	−.25	.24	0.48	−.10	.62	0.28	.12	.58	0.29
Experiment 3: Consistent non-aphantasics	−.30	.09	0.88	.18	.31	0.35	.04	.83	0.22
Experiment 3: Inconsistent responders	.26	.04*	1.27	−.28	.03*	1.72	−.14	.28	0.28

indicates significance at .05.

Table 2.17: Experiments 1 and 2: correlations between post-minus-pre instruction associative recognition d' and visual imagery measures.

	VVIQ ratings			PFT accuracy			PFT response time		
	r	p	BF_{10}	r	p	BF_{10}	r	p	BF_{10}
Experiment 1: Imagery group	−.35	.008*	5.08	.11	.44	0.22	.12	.38	0.24
Experiment 1: Control group	.10	.46	0.22	.09	.49	0.21	.03	.80	0.17
Experiment 2: Top-bottom imagery	.11	.37	0.22		N/A			N/A	
Experiment 2: Actor-object imagery	.05	.71	0.18		N/A			N/A	
Experiment 2: Standard interactive imagery	−.07	.54	0.18		N/A			N/A	

indicates significance at .05.

Table 2.18: Experiments 1 and 2: correlations between post-minus-pre instruction order recognition d' and visual imagery measures.

	VVIQ ratings			PFT accuracy			PFT response time		
	r	p	BF_{10}	r	p	BF_{10}	r	p	BF_{10}
Experiment 1: Imagery group	−.12	.36	0.25	−.03	.85	0.17	.26	.053	1.02
Experiment 1: Control group	.25	.064	0.89	.03	.84	0.17	.11	.40	0.23
Experiment 2: Top-bottom imagery	.007	.96	0.15		N/A			N/A	
Experiment 2: Actor-object imagery	.16	.16	0.38		N/A			N/A	
Experiment 2: Standard interactive imagery	.06	.60	0.23		N/A			N/A	

indicates significance at .05.

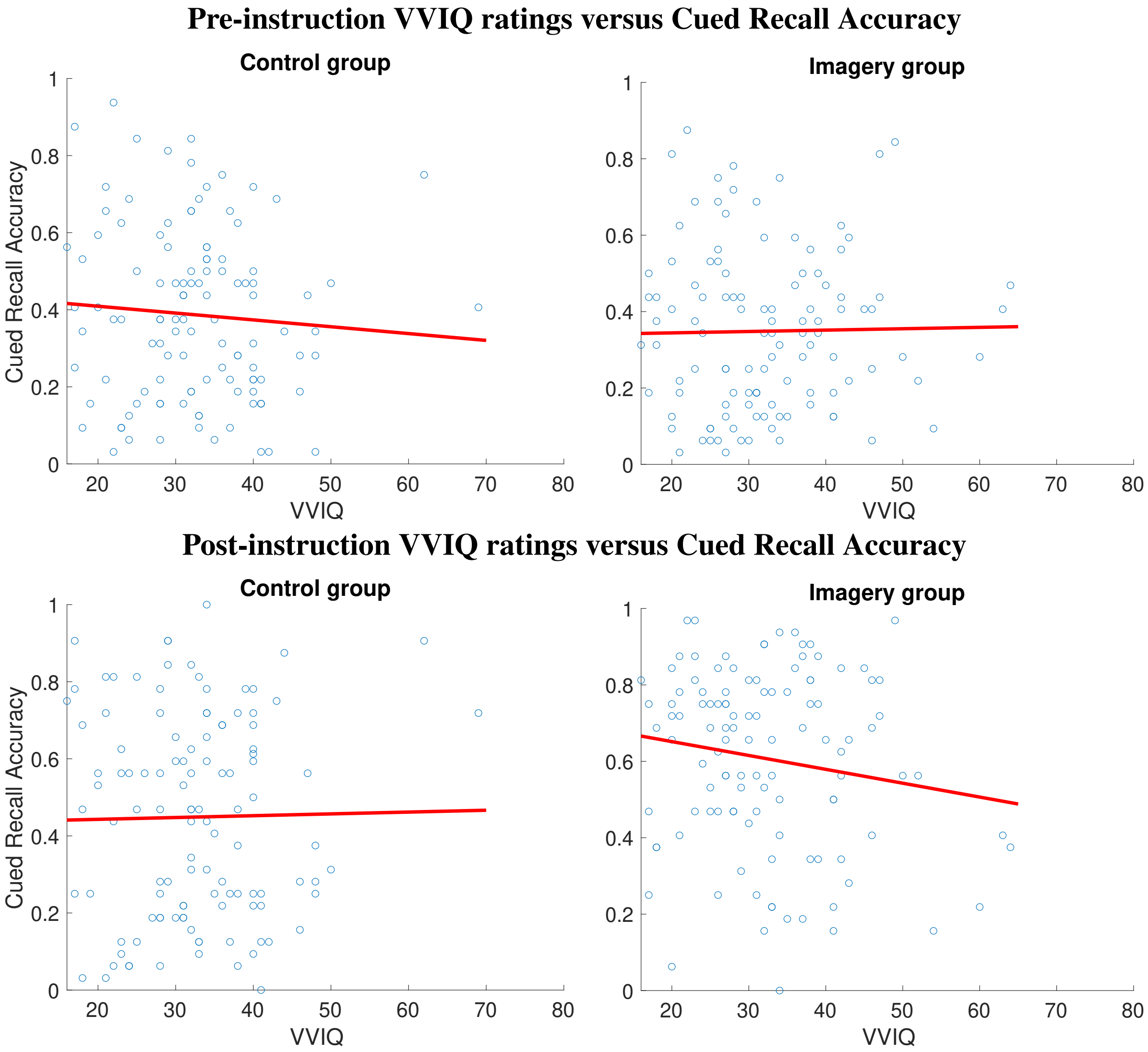


Figure 2.5: Experiment 1: Scatter plots of VVIQ ratings versus cued recall accuracy for the pre- and post-instruction phases in both imagery and control groups. Each point represents one participant. Regression lines are plotted in red.

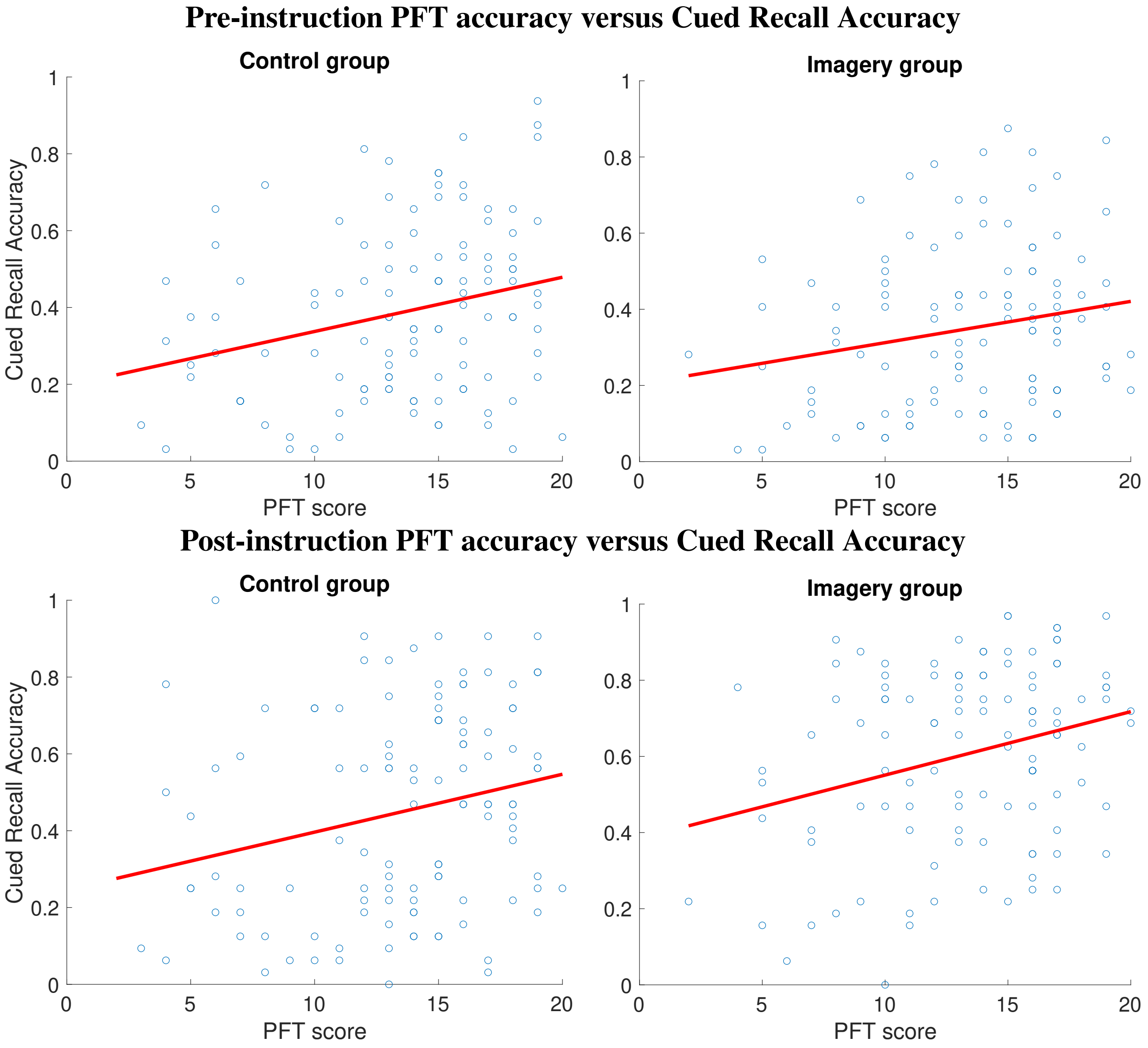


Figure 2.6: Experiment 1: Scatter plots of PFT accuracy versus cued recall accuracy for the pre- and post-instruction phases in both imagery and control groups. Each point represents one participant. Regression lines are plotted in red.

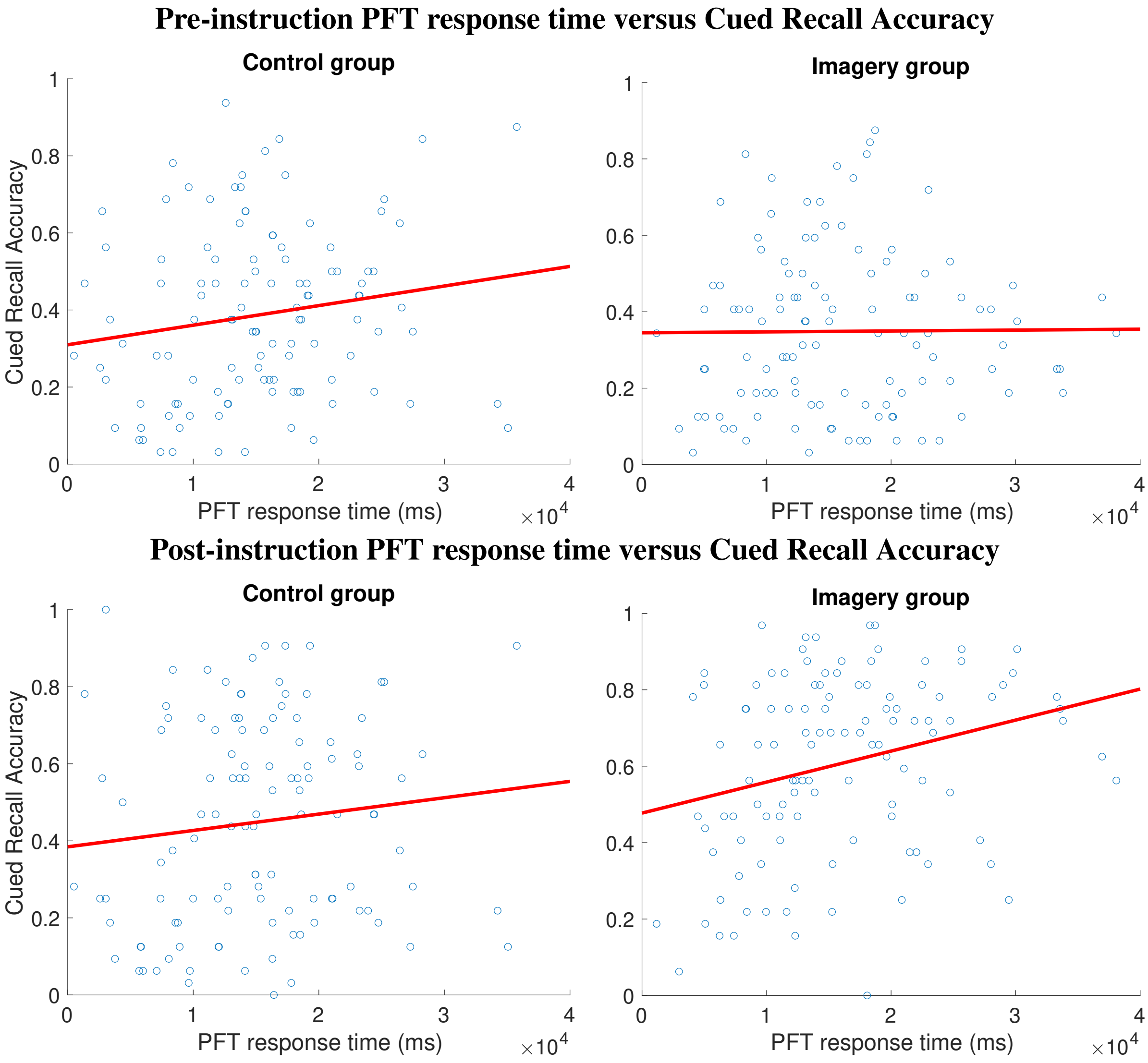


Figure 2.7: Experiment 1: Scatter plots of PFT response time versus cued recall accuracy for the pre and post-instruction phases in both imagery and control groups. Each point represents one participant. Regression lines are plotted in red.

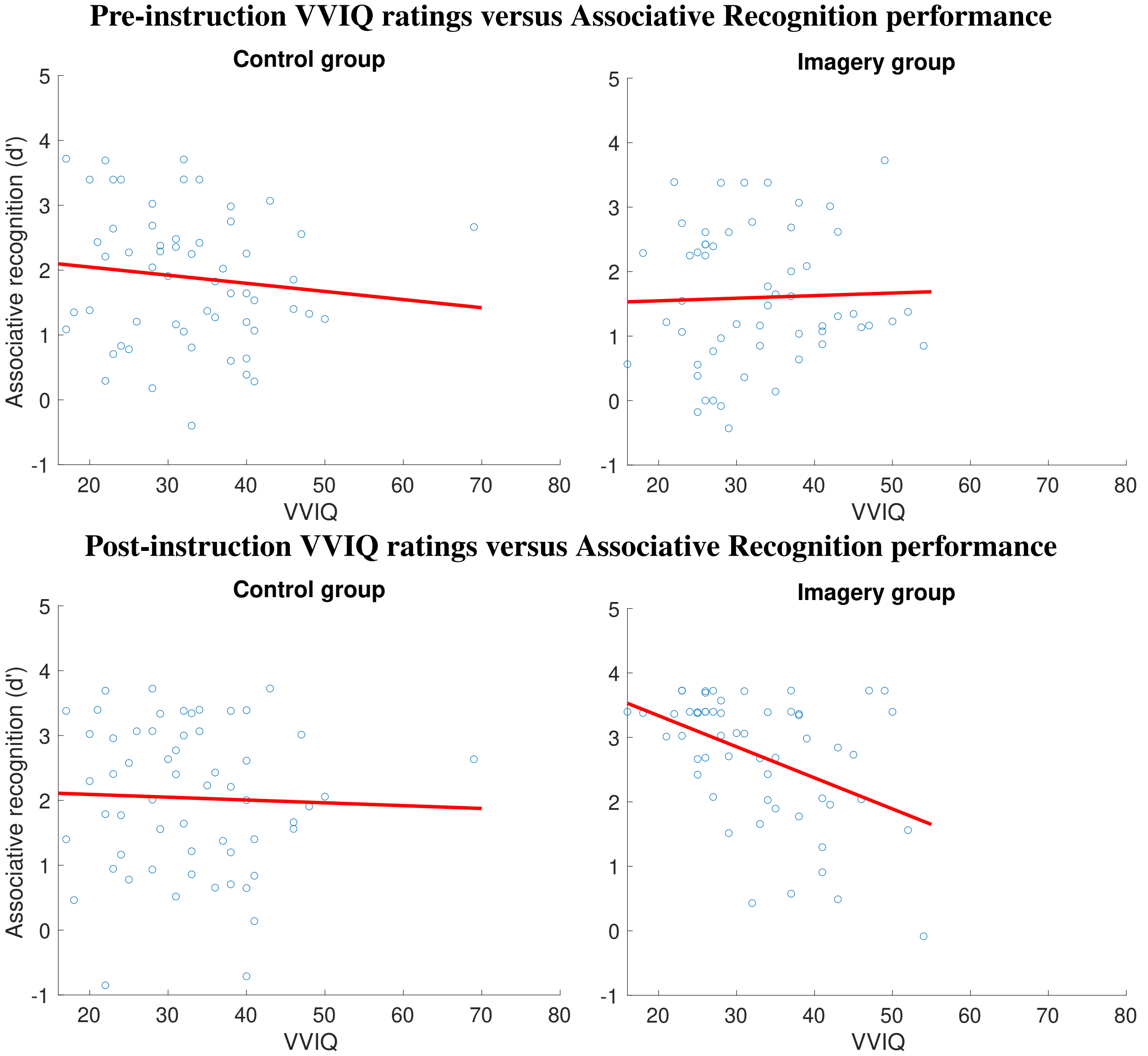
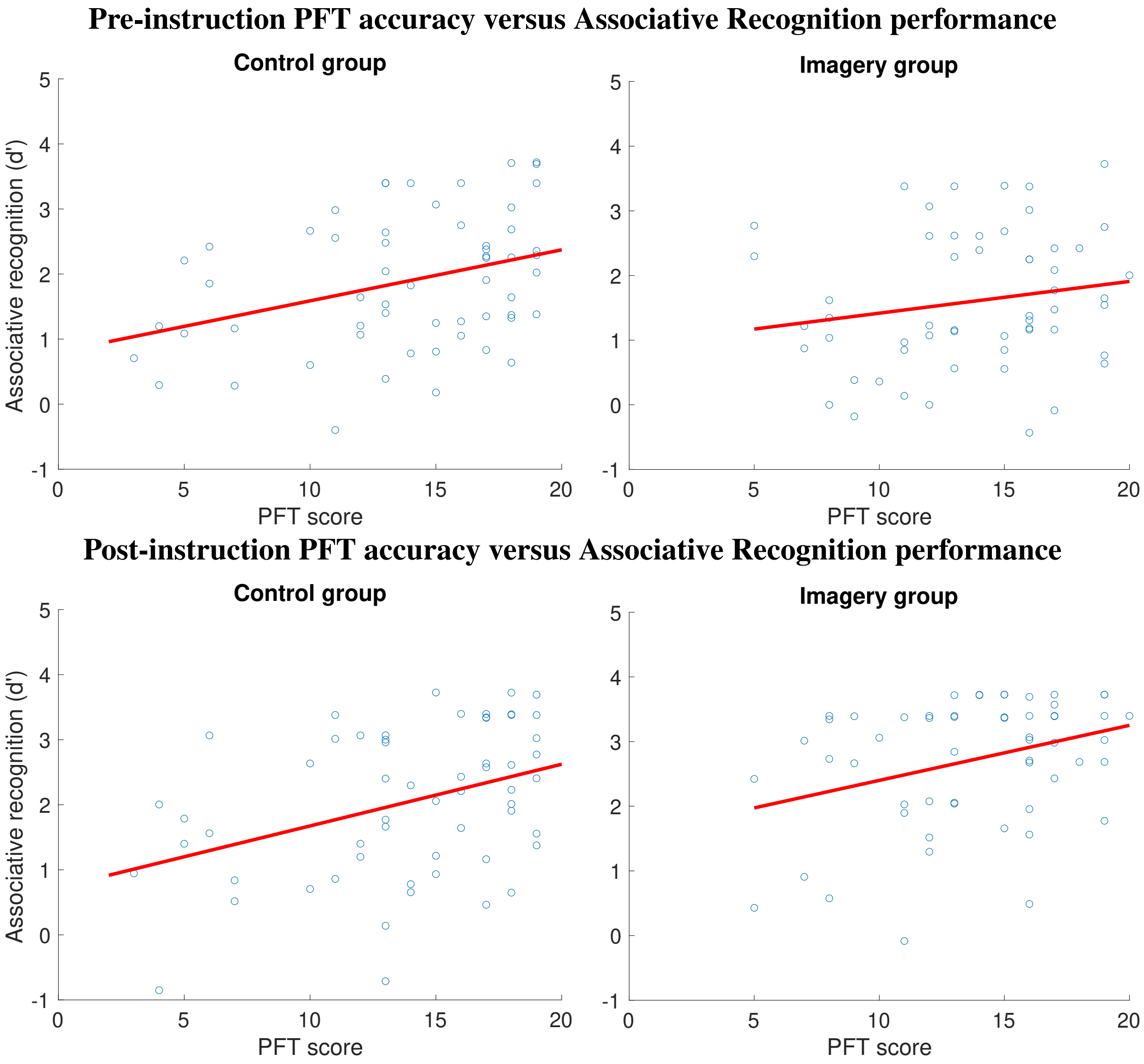


Figure 2.8: Experiment 1: Scatter plots of VVIQ ratings versus associative recognition d' for the pre- and post-instruction phases in both imagery and control groups. Each point represents one participant. Regression lines are plotted in red.


Pre-instruction PFT accuracy versus Associative Recognition performance

Control group

Imagery group

Post-instruction PFT accuracy versus Associative Recognition performance

Control group

Imagery group

Figure 2.9: Experiment 1: Scatter plots of PFT accuracy versus associative recognition d' for the pre- and post-instruction phases in both imagery and control groups. Each point represents one participant. Regression lines are plotted in red.

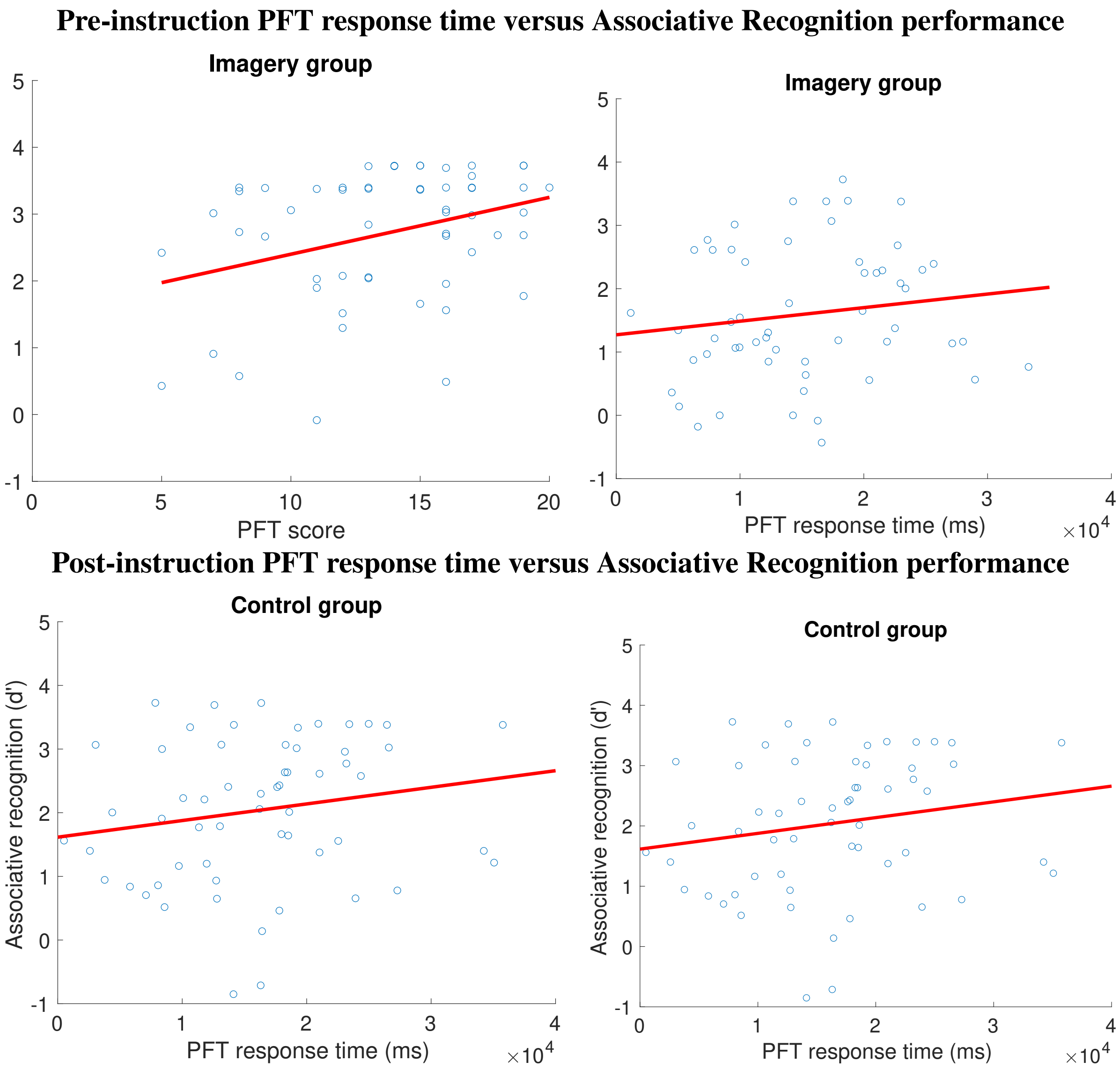


Figure 2.10: Experiment 1: Scatter plots of PFT response time versus associative recognition d' for the pre- and post-instruction phases in both imagery and control groups. Each point represents one participant. Regression lines are plotted in red.

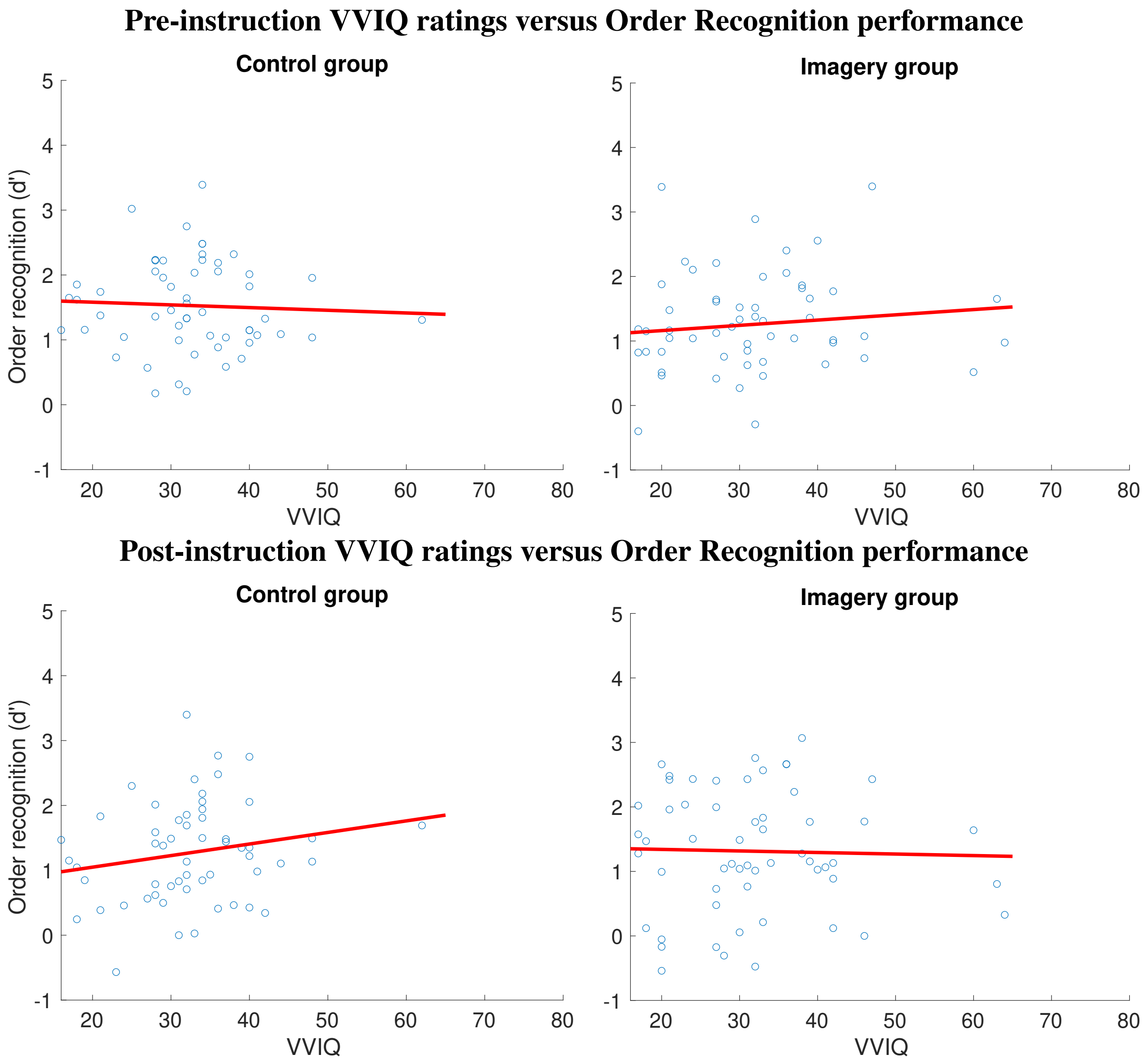


Figure 2.11: Experiment 1: Scatter plots of VVIQ ratings versus order recognition d' for the pre- and post-instruction phases in both imagery and control groups. Each point represents one participant. Regression lines are plotted in red.

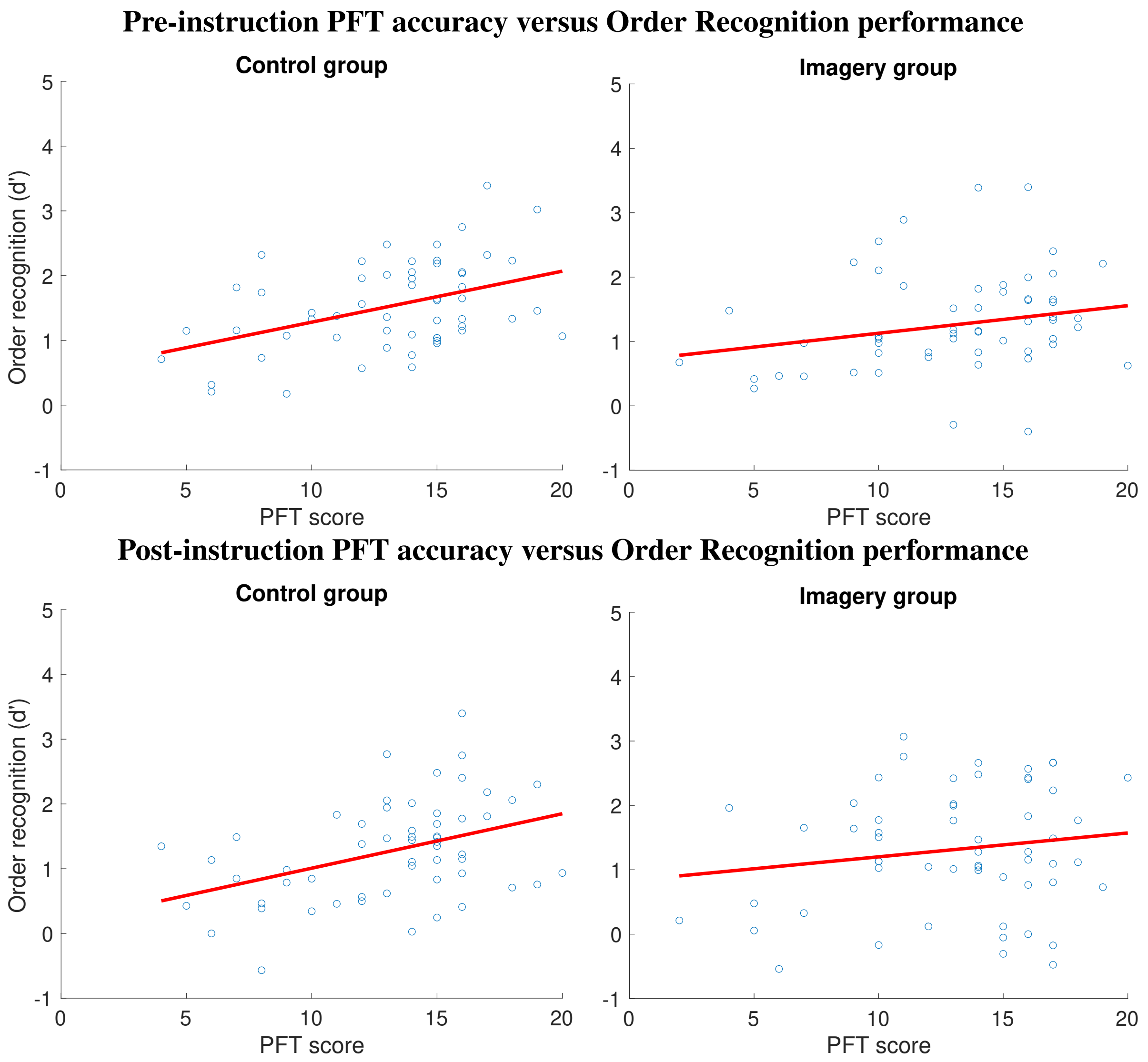


Figure 2.12: Experiment 1: Scatter plots of PFT accuracy versus order recognition d' for the pre- and post-instruction phases in both imagery and control groups. Each point represents one participant. Regression lines are plotted in red.

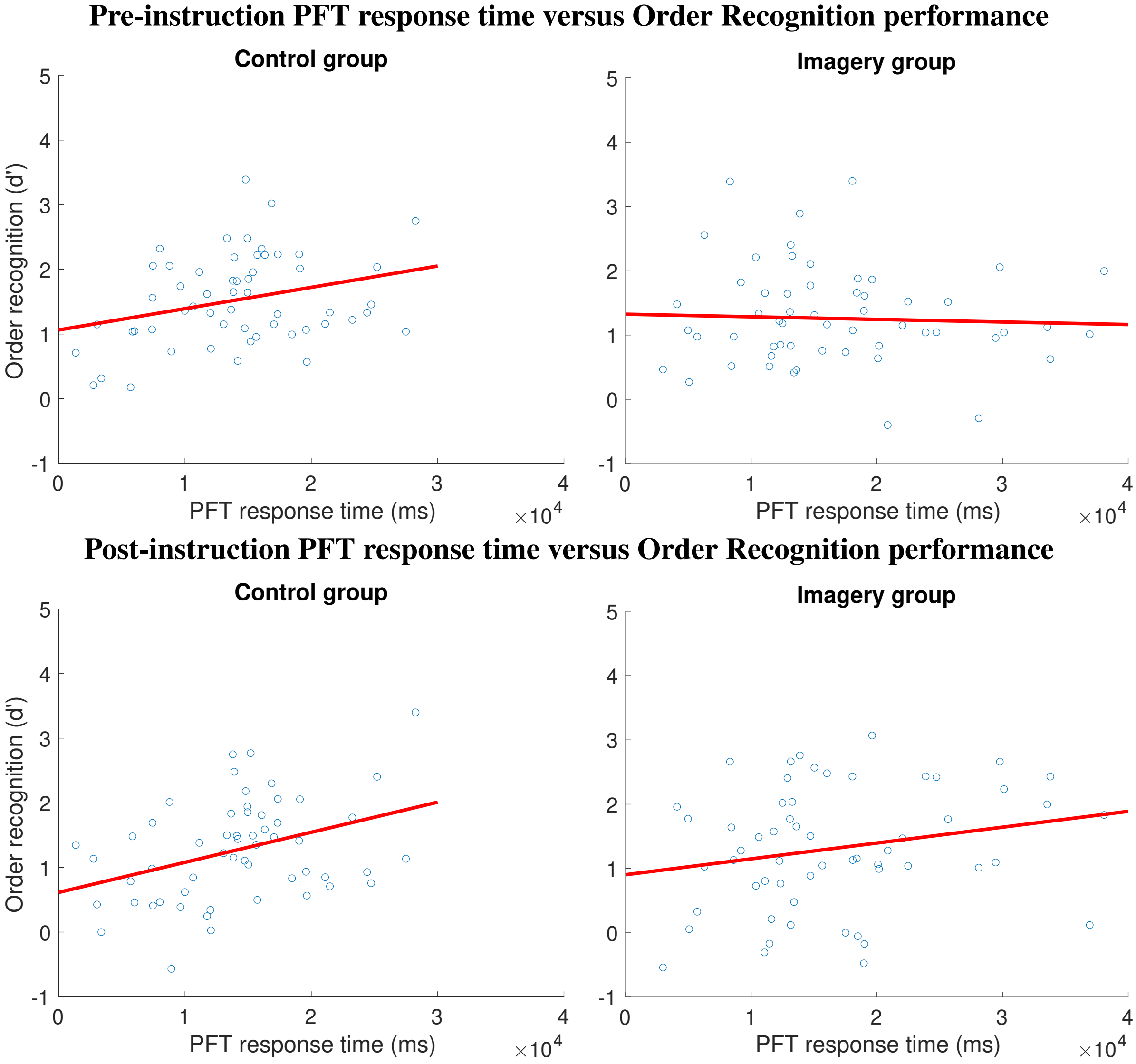


Figure 2.13: Experiment 1: Scatter plots of PFT response time versus order recognition d' for the pre- and post-instruction phases in both imagery and control groups. Each point represents one participant. Regression lines are plotted in red.

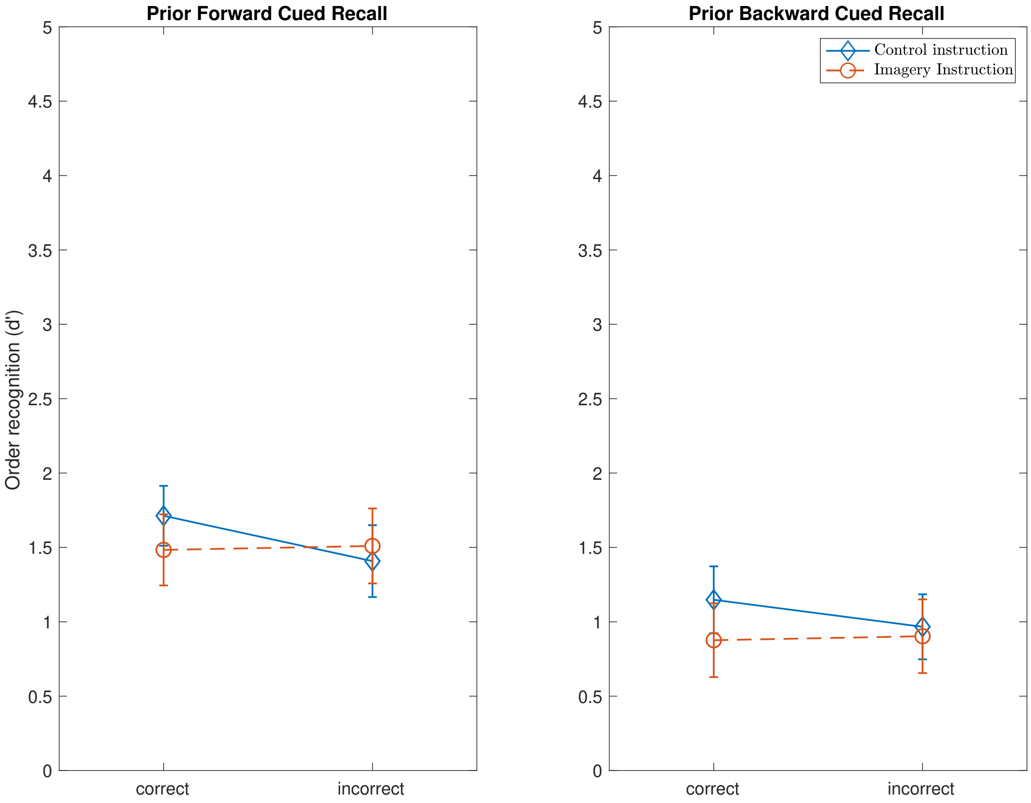


Figure 2.14: Experiment 1: Order recognition performance for pairs tested with forward cued recall and for pairs tested with backward cued recall. Error bars represent 95% confidence intervals based on standard error of the mean.

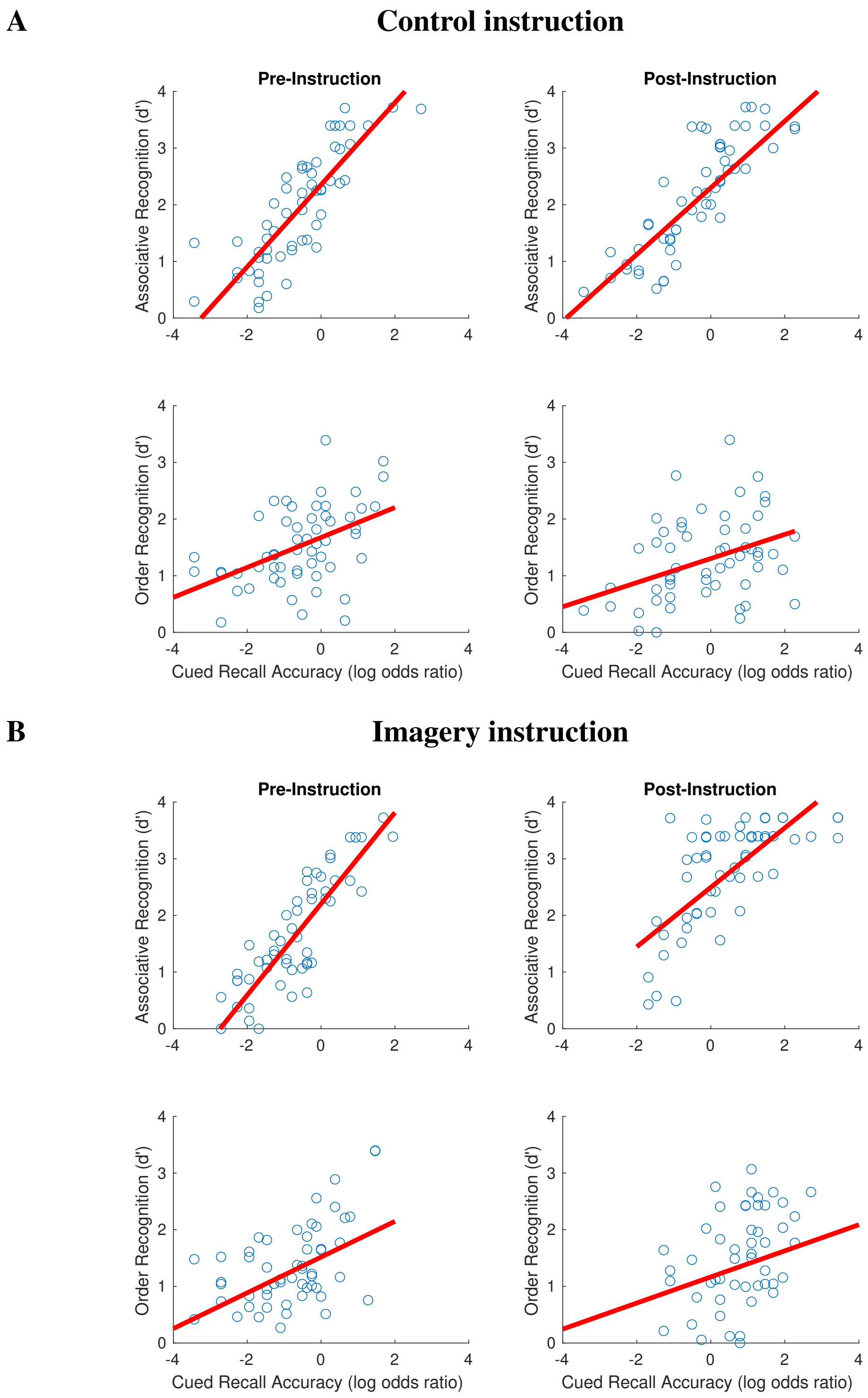


Figure 2.15: Experiment 1: Scatter plots of log-odds transformed cued recall accuracy versus associative recognition performance, and versus order recognition. Regression lines are plotted in red. This measured the relationship between both associative and order recognition to cued recall accuracy. Each point is a single participant.

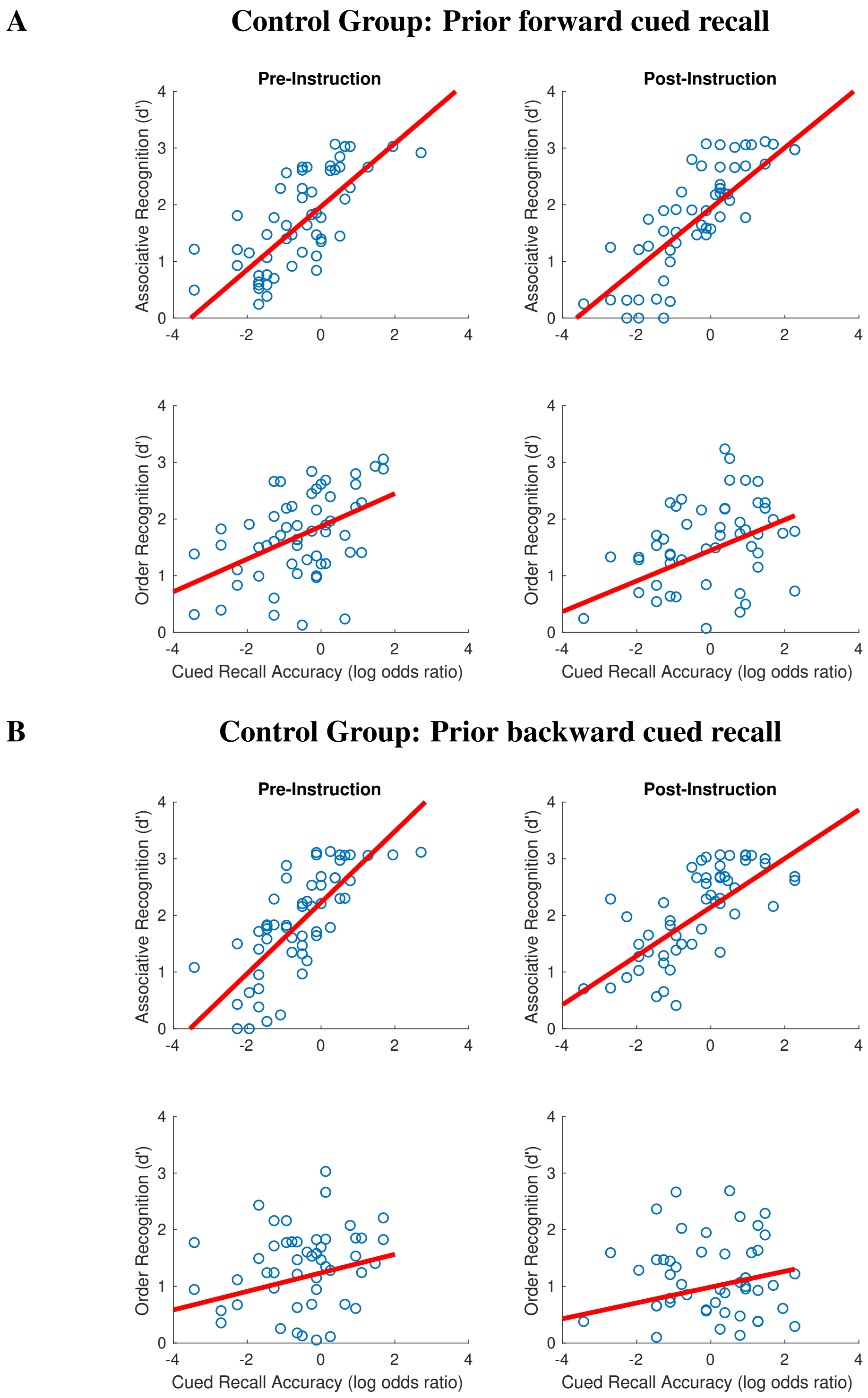


Figure 2.16: Experiment 1, Control group: Scatter plots of control group log-odds transformed cued recall accuracy versus associative and order recognition for (Top) pairs tested with forward cued recall (Bottom) pairs tested with backward cued recall. Regression lines are plotted in red. Each point is a single participant.

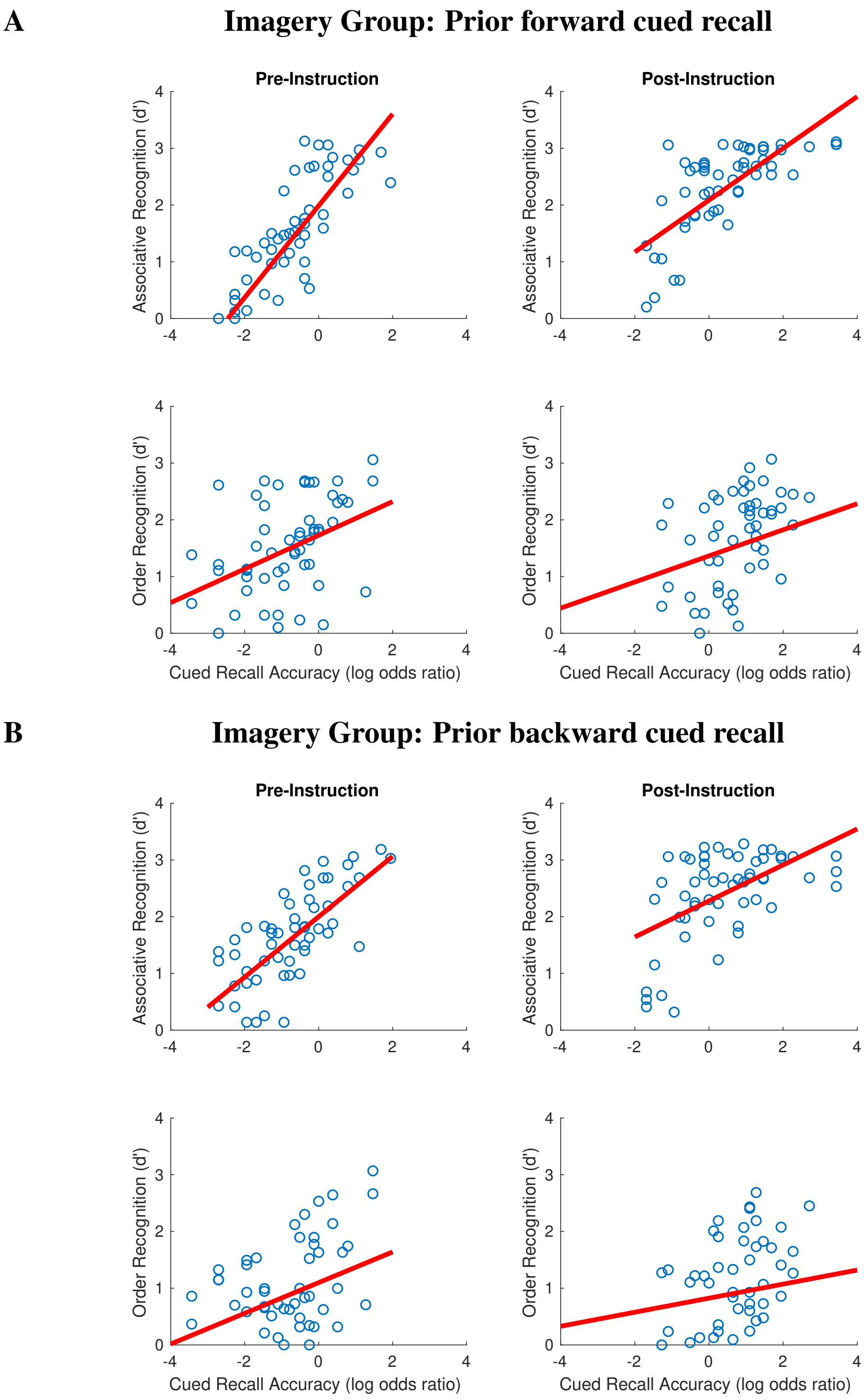


Figure 2.17: Experiment 1, Imagery group: Scatter plots of imagery group log-odds transformed cued recall accuracy versus associative and order recognition for (Top) pairs tested with forward cued recall (Bottom) pairs tested with backward cued recall. Regression lines are plotted in red. Each point is a single participant.

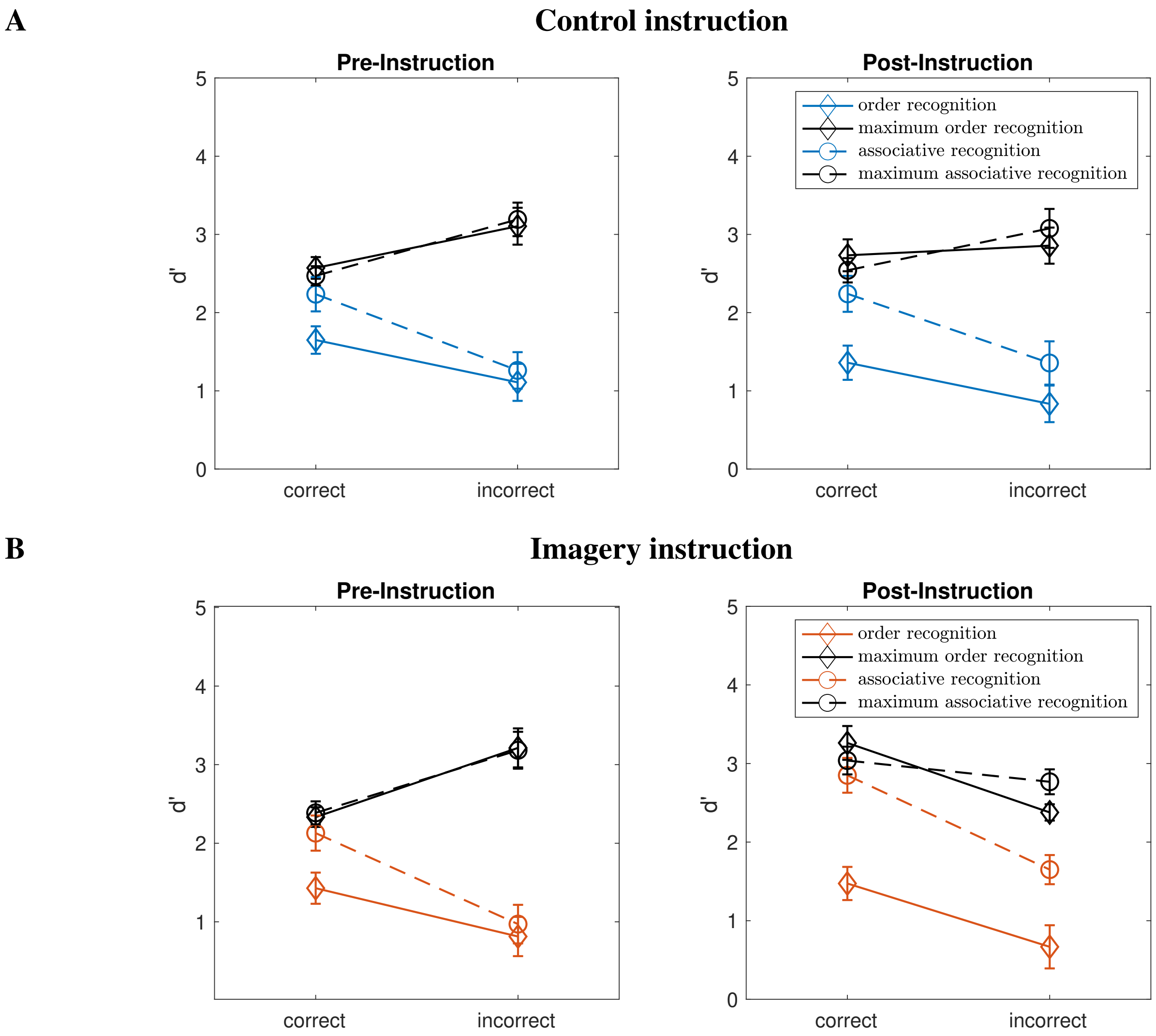


Figure 2.18: Experiment 1: Associative and order recognition performance from experiment 1 computed separately for correctly versus incorrectly recalled pairs. Also plotted is d'_{max} for each measure (see methods). This measured the within-subject relationship between both order and associative recognition to cued recall performance. Error bars represent 95% confidence intervals based on standard error of the mean.

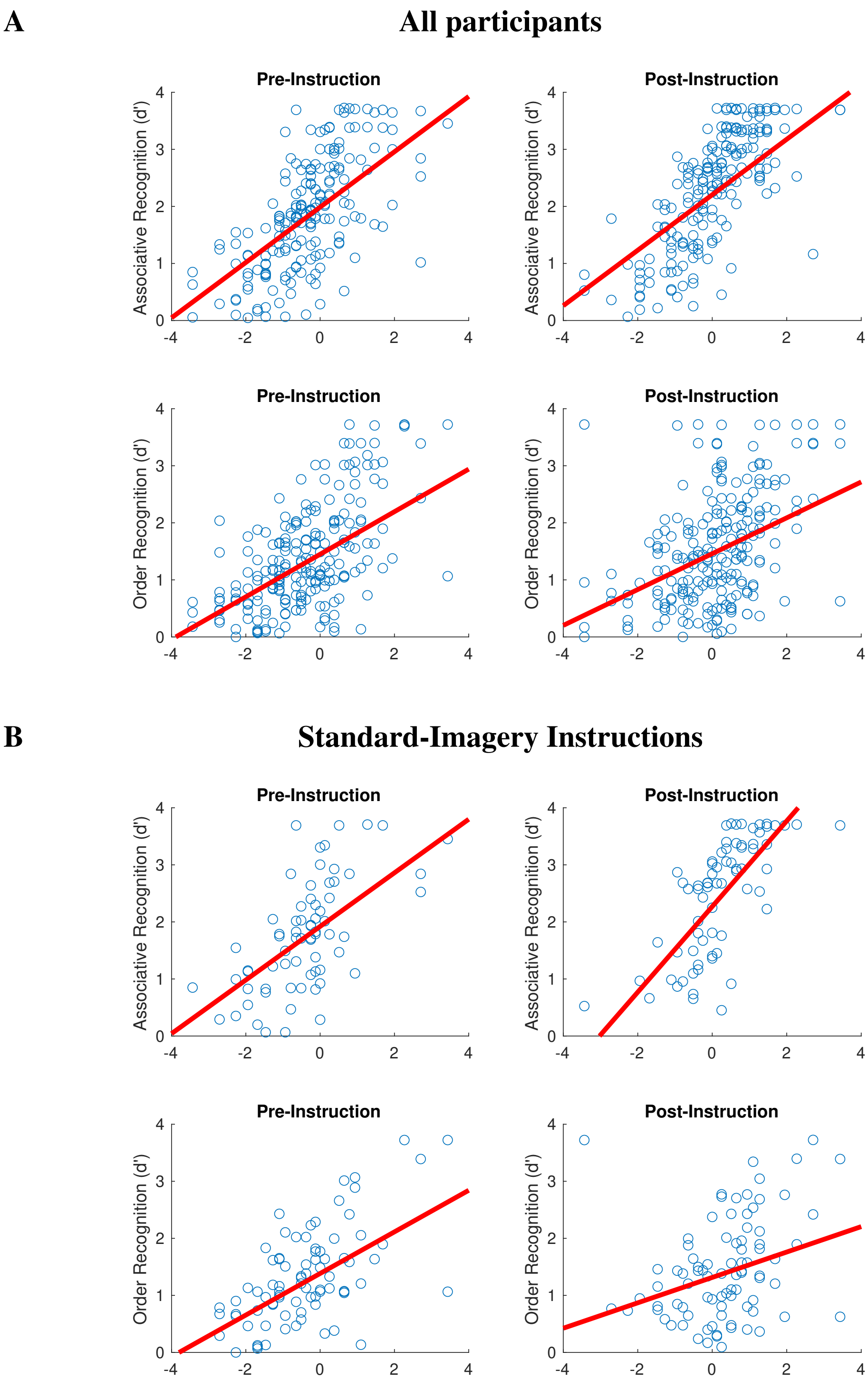


Figure 2.19: Experiment 2: Scatter plots of log-odds transformed cued recall accuracy versus associative recognition performance, and versus order recognition. Regression lines are plotted in red. (Top) Scatter-plots for all participants, collapsed across groups. (Bottom) Scatter-plots for the standard-imagery group. This measured the between-subject relationship between both associative and order recognition to cued recall accuracy. Each point is a single participant.

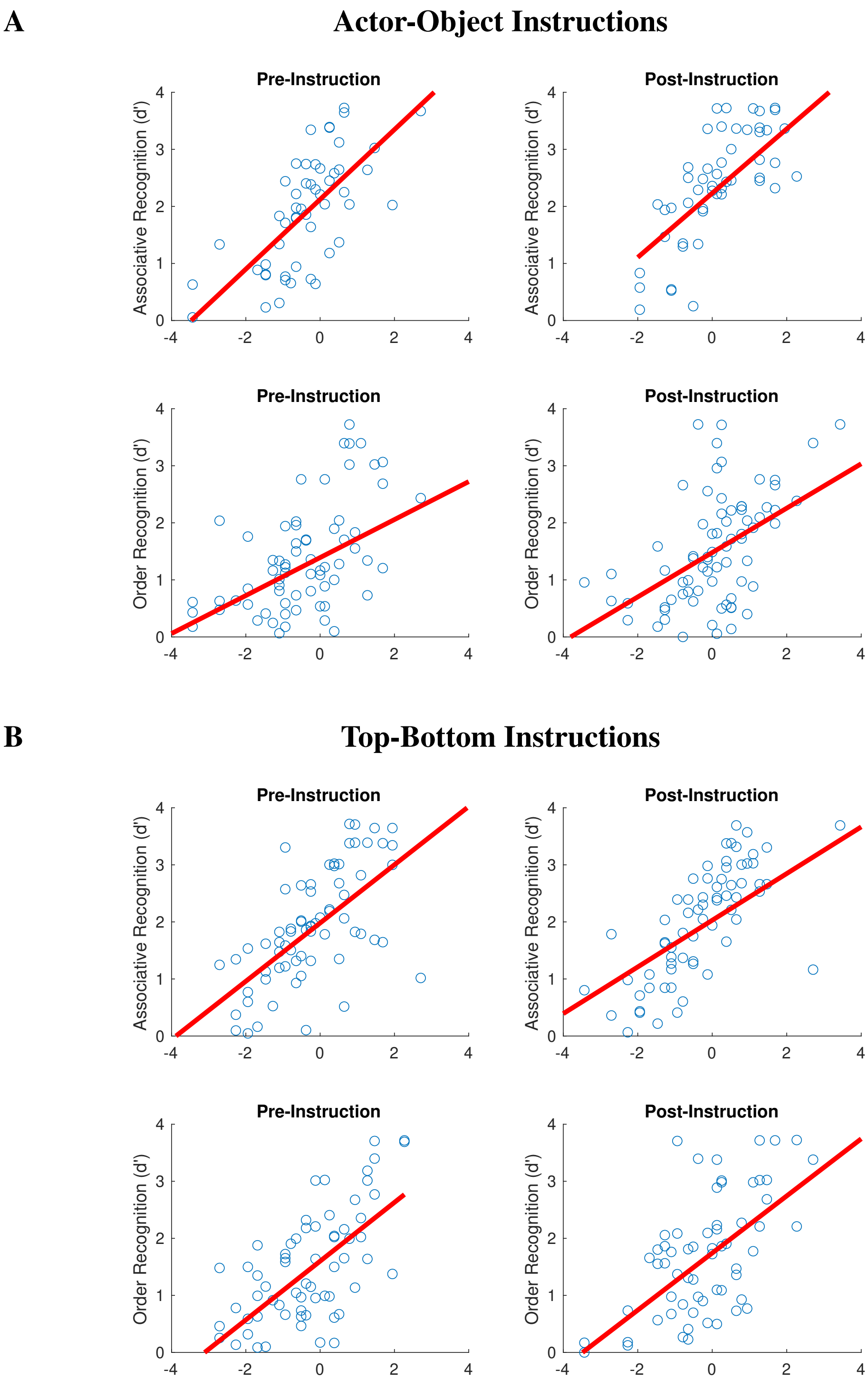


Figure 2.20: Experiment 2: Scatter plots of log-odds transformed cued recall accuracy versus associative recognition performance, and versus order recognition. Regression lines are plotted in red. (Top) Scatter-plots for the actor-object imagery group. (Bottom) Scatter-plots for the top-bottom imagery group. This measured the relationship between both associative and order recognition to cued recall accuracy. Each point is a single participant.

Chapter 3

Modelling constituent-order despite symmetric associations in memory

Abstract

Mathematical models of association memory either predict that knowledge for constituent order of a word pair (AB vs. BA) is perfectly unrelated, or completely dependent on knowledge for the pairing itself. Behavioural data contradicts both predictions; when a word pair can be remembered (given A, recall B), knowledge for its constituent-order is above chance, but still fairly low. The inherent symmetry of convolution has enabled convolution-based models to explain symmetry of associative strengths, but offers no way to discriminate AB from BA. We evaluated four ways to extend convolution to store order, where order is encoded as item features, partial permutations of features, explicit item-position associations, or the addition of item and position vectors before convolution. All approaches were successful in discriminating order within behaviourally observed ranges, without compromising associative symmetry. Only the permutation model passed a further challenge, disambiguate AB from BC in double-function lists, as humans can do. It is possible that each of our proposed mechanisms might apply to a different, particular task setting. However, the partial permutation model can thus far explain the broadest set of empirical benchmarks.

3.1 Introduction

Memory for associations forms the cognitive basis for a large portion of behaviour (Murdock, 1974; Lashley, 1951). In many cases, such as remembering face-name relationships at a dinner party, or that colorful snakes are poisonous, it is sufficient to remember that stimuli are associated to each other. But sometimes it is important to remember an association along with its constituent-order (AB versus BA). Indeed, many examples of order-sensitive associations exist in language, such as modifier-head relationships in compound words, PAN CAKE versus CAKE PAN, or HOUSE GUEST versus GUEST HOUSE (Dressler, 2006; Caplan, Boulton, & Gagné, 2014). However, memory for order has typically not been a focus in the experimental study of verbal association memory. Standard tests of association memory ask participants to study pairs of words (AB), followed by cued recall (given A, respond with B). Participants can respond with B when given A, and vice versa, without knowing the constituent-order of the pairing. Moreover, memory for order is typically studied with separate tasks such as serial recall (study A, B, C, D, recall the list in order).

Consequently, mathematical models of association memory are quite poor at accounting for constituent-order, either assuming that associations are stored with perfect order, or with no order at all. Models based on convolution (Kelly et al., 2013; Murdock, 1982; Metcalfe Eich, 1982; Plate, 1995), and recent models within the REM framework (Cox & Criss, 2017, 2020; Criss & Shiffrin, 2005), assume associations are stored with no order. In these models, AB is mathematically equivalent to BA. The face-value prediction is that memory for constituent-order will be at chance. However, given evidence that participants can remember constituent-order above-chance (Greene & Tussing, 2001; Kato & Caplan, 2017; Kounios et al., 2001, 2003; J. Yang et al., 2013), one might rescue convolution, and other symmetric models, by allowing for some additional source of information to support order judgments, such as an additional term in the model. The consequence of storing order separately from associations is that the models would predict that memory for constituent-order should be unrelated to memory for the pairing itself. The second type of prediction, that associations are stored with perfect order, comes from matrix models (Anderson, 1970; Humphreys, Bain, & Pike, 1989; Pike, 1984) and models that concatenate the two item vectors (Hintzman, 1984; Shiffrin & Steyvers, 1997). These models can infer order with no ambiguity, predicting that memory for constituent-order (AB versus BA) should be perfect given that the association itself can be recalled.

Kato and Caplan (2017) tested these predictions with a task we refer to as order recognition (Greene & Tussing, 2001; Kounios et al., 2001, 2003; J. Yang et al., 2013). Order recognition tests memory for constituent-order directly by presenting pairs in their original (AB) or reversed order (BA). Participants then provide a forced-choice judgment whether the probe is intact or reverse. One group of participants were tested with cued recall, and then order recognition for each studied pair, and compared to another group tested with associative recognition after cued recall instead.[1] Matrix models predict that order recognition performance should be perfect for correctly recalled pairs. Convolution models predict that order recognition performance should be equiv-

[1] Cued recall and associative recognition should be highly correlated because they essentially test the same information— knowledge of pairings between words. Thus, the cued recall-associative recognition group provided a realistic upper ceiling to compare the order recognition group against.

alent for correct and incorrectly recalled pairs. Contradicting both predictions, order recognition was significantly better when cued recall was correct, but well below maximum, and well below associative recognition for correctly recalled pairs.[2] These results indicate that verbal associations are neither encoded with perfect order, nor are completely order-absent, inconsistent with assumptions in all models.

Another clue about the representation of associations and their constituent order comes from from double function lists in Rehani and Caplan (2011), where cued recall was direction-specific. Double function lists (Howard, Jing, Rao, Provyn, & Datey, 2009; Primoff, 1938; Rehani & Caplan, 2011; Slamecka, 1976), contain pairs, where each constituent item appears in two pairs, once in the left position, and once in the right position (AB, . . . , BC, . . . , CA, . . .). Consider that B is presented as a cue on the left-hand side. Correctly responding with C requires knowledge of relative position/order, for example, that B appeared on the left in pair BC, but not AB. Performance is compared to single function pairs that do not share items (EF, . . . , GH, . . . , IJ, . . .). Because of their extreme assumptions about order, matrix and convolution models generate direct predictions about this task. A convolution model has no information to select between A and C. Thus, assuming the model guesses between two possible responses, convolution predicts cued recall accuracy for double function pairs will be one-half that of single-function pairs. In contrast, matrix-based models suffer no interference between AB and BC (see below). Therefore, the model predicts equal accuracy for double and single-function pairs. Contradicting both matrix and convolution model predictions, Rehani and Caplan (2011) found double-function cued recall accuracy was somewhat lower, but well above one-half of single-function accuracy, converging with evidence from the order recognition task that associations are neither stored order-absent, nor with perfect directionality.[3]

In sum, participants can discriminate AB versus BA during a word pair task (Greene & Tussing, 2001; Kato & Caplan, 2017; Kounios et al., 2001, 2003; J. Yang et al., 2013), and even use order/item-position information to aid cued recall (B ?) to solve AB versus BC interference (Rehani

[2] Kato and Caplan (2017) also addressed the possibility that testing with cued recall influenced order recognition. In their second experiment they withheld half the pairs from cued recall testing, and in their third experiment moved cued recall to the end of the session. In both cases they found that the order-cued recall relationship persisted.

[3] One could argue that the ability to disambiguate double function pairs does not come from memory for order, but rather, because each item in these pairs was repeated, and thus more available in memory. However, Caplan, Rehani, and Andrews (2014) found when participants were able to respond with both associates for double function pairs, double and single function cued recall accuracy was equivalent, arguing against this confound.

& Caplan, 2011). Taken together, this suggests that the constituent-order of verbal associations is explicitly stored, and in a way that is moderately dependent on memory for the pairing itself.

3.1.1 Associative Symmetry

Despite evidence that associations are stored with moderate levels of order, there is also a sense in which verbal associations are rather symmetric. Initial support for idea, known as associative symmetry, arose from the stable tendency for forward cued recall accuracy (APPLE ?) and backward cued recall (? OVEN) accuracy to be equal on average (Asch & Ebenholtz, 1962; Horowitz, Brown, & Weissbluth, 1964; Kahana, 2002; Kato & Caplan, 2017; Murdock, 1962). However, Kahana (2002) showed that an asymmetric model could produce symmetry in mean cued recall accuracy, suggesting this result is not diagnostic of symmetric associations. Instead, Kahana (2002) proposed that associative symmetry should be tested at the pair level, with two cued recall trials for each word, and where test 1 and test 2 is either forward or backward cued recall. Indeed, multiple studies have returned a near-perfect correlation for incongruent conditions (forward-backward, backward-forward), that are remarkably close to what are essentially test-retest correlations for congruent conditions (forward-forward, backward-backward) (Kahana, 2002; Kato & Caplan, 2017; Rehani & Caplan, 2011; Rizzuto & Kahana, 2000, 2001; Sommer, Schoell, & Büchel, 2008). These findings either suggest forward and backward cued recall are testing the same bi-directional association in memory, or, that there are distinct forward and backward associations for a given pair, but these are highly correlated in their strengths (Kahana, 2002).

We were particularly interested in associative symmetry here because of the potential paradox between association memory that is highly symmetric, yet supports memory for its constituent-order. As we elaborate below, it was especially challenging in previous attempts to modify matrix models to simultaneously produce moderate order memory and associative symmetry (Kato & Caplan, 2017). A strong account of association memory should be able to account for both constraints, and thus, we include this as an additional benchmark for all models.

3.1.2 Attempts to produce order and symmetry in current models

Next, we describe how current models fare against the constraints of associative symmetry and moderate memory for order.

Matrix-based models Associations are encoded as follows, $M = \mathbf{a}\mathbf{b}^\intercal$, where M denotes the memory matrix, **a** and **b** represent item vectors, and $\intercal$ denotes transpose. Bold-face indicates column vectors. Cued recall is simulated with matrix multiplication, for example, $M\mathbf{b} \approx \mathbf{a} + noise$. Matrix multiplication is direction sensitive, meaning that $\mathbf{b}^\intercal M \approx 0 + noise$. By comparing the outputs of $M\mathbf{b}$ and $\mathbf{b}^\intercal M$, the model can unambiguously infer that item **b** appeared in the left position. For similar reasons, matrix models also have a perfect ability to solve double function interference. If two pairs that share an item are stored in memory, $M = \mathbf{a}\mathbf{b}^\intercal + \mathbf{b}\mathbf{c}^\intercal$, the direction specificity of forward and backward cued recall means that a given item vector **b** can cue completely different pairs in memory based on direction, $M\mathbf{b} \approx \mathbf{a}$ and $\mathbf{b}^\intercal M \approx \mathbf{c}$.

One can eliminate this directionality by simultaneously storing the forward and reverse association, $\alpha_f \mathbf{a}^\intercal \mathbf{b} + \alpha_b \mathbf{b}^\intercal \mathbf{a}$, where α_f and α_b are scalar random values that represent variable encoding strengths. Assuming that α_f and α_b are perfectly correlated, and that $\mathrm{E}\left[\alpha_f\right] = \mathrm{E}\left[\alpha_b\right]$, this model could produce perfect associative symmetry (Kahana, 2002), but as a direct consequence, cannot discriminate AB from BA (Kato & Caplan, 2017) or solve double function lists (Rehani & Caplan, 2011). To regain some ability to disambiguate AB from BA, $\mathrm{E}\left[\alpha_f\right]$ could be increased relative to $\mathrm{E}\left[\alpha_b\right]$, so that the forward association is stronger in memory; however, the model now produces an forward recall advantage violating associative symmetry, and predicts order recognition performance would positively correlate with the difference between forward and backward cued recall performance. Kato and Caplan (2017) found no evidence for the latter prediction, these correlations were not significant. Kato and Caplan (2017) also tested a model that sometimes encoded pairs in the incorrect order with probability p_{rev}. Increasing p_{rev} reduced the model's order recognition performance, even to the moderate levels seen in behaviour. However, the model assumes that even wrong order judgments are made with perfect certainty, because they come from perfectly directional associations in memory. The resulting prediction is that participants should be unlikely to switch their response if they are tested twice for order recognition, correct-correct or incorrect-incorrect judgments should be most frequent. This prediction was also unsupported in Kato and Caplan's (2017) data—participants did not stick with their order judgments more frequently than they switched their order judgments. Along with evidence from other analyses, order judgments seem to not be made with perfect certainty, but are rather more like uncertain, noisy decisions that are prone to change on retest.

Convolution-based models Convolution models do not store order at all. Associations are stored as follows, $\mathbf{m} = \mathbf{x} * \mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ denote item/word vectors, $\mathbf{m}$ denotes the memory vector, and $*$ denotes circular convolution. Importantly, convolution is strictly commutative, $\mathbf{a} * \mathbf{b} \equiv \mathbf{b} * \mathbf{a}$. This property causes convolution to naturally produce associative symmetry (Kahana, 2002), but also means that there is no way to recover the constituent-order of the pair after encoding. To retain order information in a convolution model, one could permute the elements of item-vectors before encoding (Jones & Mewhort, 2007; Kelly et al., 2013; Plate, 1995; Recchia, Jones, Sahlgren, & Kanerva, 2010), expressed as follows, $\mathbf{m} = p_l(\mathbf{x}) * p_r(\mathbf{y})$, where p denotes permutation operator, and subscript l and r indicate the position-specific permutation pattern applied to each vector. Permutation allows convolution to encode order-sensitive relationships (Jones & Mewhort, 2007), along with other useful side-effects (Kelly et al., 2013); however, in published implementations, the whole vector has been permuted, which effectively implements a non-commutative operation, more like a matrix-outer product, $p_l(\mathbf{x}) * p_r(\mathbf{y}) \neq p_r(\mathbf{x}) * p_l(\mathbf{y})$. Thus, fully permuting item vectors may be incompatible with empirical data in a similar way as an unmodified matrix model, although we do test this idea, with a small twist, below.

3.1.3 Extending convolution to store order

In sum, the concurrent empirical constraints of associative symmetry and moderate order memory prove difficult for all existing models. Convolution models and modified matrix models can produce perfect associative symmetry, but disregard order, while non-commutative versions of both matrix and convolution models over-predict the degree to which order is remembered. One could address these challenges with two possible approaches, either modify matrix models to have reduced order memory, or extend convolution models to store order. In the present article we take the latter approach. Our objective here is not to fundamentally alter basic model mechanisms, but to present minor modifications that could allow convolution to produce mid-range order memory while preserving useful characteristics, like associative symmetry, that make convolution a rich account of verbal association memory.

To this end, we designed four simple modifications (Illustrated in Figure 3.1) that could encode order without significantly increasing model complexity,

- **Model A** (Figure 3.1a): Order is encoded as explicit associations between item vectors and "position" vectors, bearing some resemblance to positional-coding models of serial recall

(Conrad, 1960; Brown, Neath, & Chater, 2007; Burgess & Hitch, 1999; Farrell, 2012; Henson, 1998), or item-context associations in the Temporal Context Model (Howard & Kahana, 1999) but with just two unique position vectors. These two associations for the left and right positions are stored along with the item-item association.

- **Model** Σ (Figure 3.1b): Similar to model A, position vectors are used to represent order, but are instead added element-wise to each item before convolving, mathematically similar to extensions of TODAM (Murdock, 1995), where item vectors were summed before convolving.

- **Model** ϕ (Figure 3.1c): Order is encoded by incorporating dedicated positional feature values into the item vector alongside item-unique features. This bears some resemblance to the ways in which numerous models have incorporated attributes such as list context as specialized features. All items in the left position receive the same set of positional feature values, and likewise for right position items.

- **Model** Π (Figure 3.1d): To encode order, item-unique feature values are shuffled or permuted in a pattern that is specific to the position of that item vector. This mechanism is directly derived from other models (Jones & Mewhort, 2007; Kelly et al., 2013; Plate, 1995), but the key difference in our implementation is that a subset of features are permuted, rather than the entire vector.

3.1.4 Summary of modelling approach

Our evaluation of these models will proceed as follows. First, we formally describe each model and its accompanying assumptions. Then, we simulate order recognition, cued recall, and associative recognition to show the relationship between model performance and key model parameters. Next, we fit models to data, to determine if each can produce a moderate relationship between order recognition and cued recall, while preserving the near-perfect correlation between forward and backward cued recall (benchmark 1a). Next, we dissect order recognition data even further, and evaluate each model against order recognition data for individual participants (benchmark 1b), and the between-subject correlations between order and associative recognition to cued recall (benchmark 1c). Finally, we evaluate each model against double function lists (benchmark 2), to test whether or not each model solve this task.

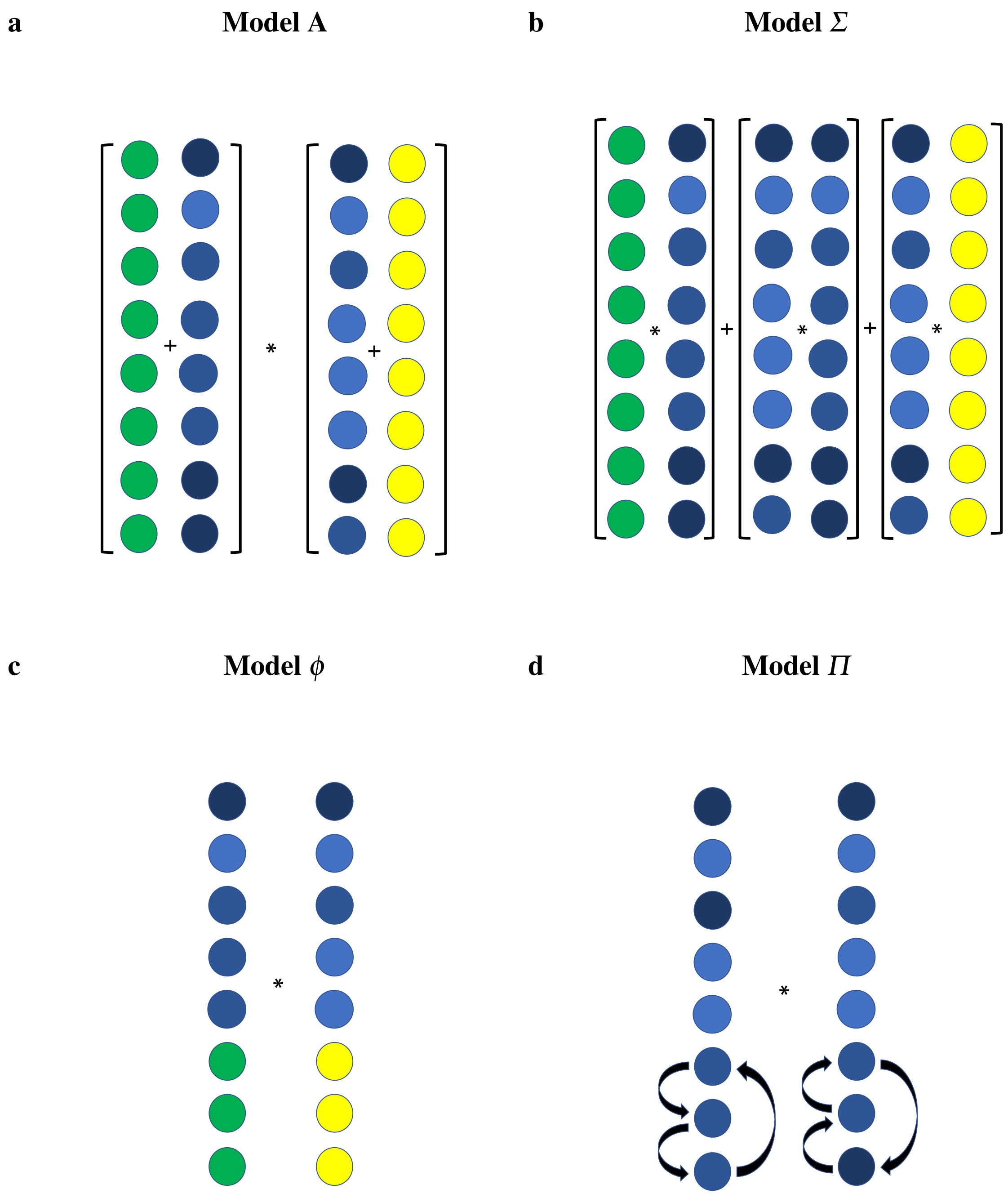


Figure 3.1: Four different mechanisms to store the constituent-order of associations within a convolution model. Blue circles denote item features, green circles correspond to the left position, and yellow circles correspond to the right position. Note, the color of the circles were for illustrative purposes, and do not indicate that features were the same value.

3.1.5 Four ways to extend convolution to store order

Before evaluating models against empirical benchmarks, we begin by formally describing each model and its accompanying assumptions.

Model A

Associations between item vectors and positional vectors are stored alongside the item-item association itself,

$$\mathbf{m}_{\mathrm{A}} = \sum_{i=1}^{L} \alpha_i \left((\mathbf{f}_i * \mathbf{l}) + (\mathbf{g}_i * \mathbf{r}) + (\mathbf{f}_i * \mathbf{g}_i) \right), \tag{3.1}$$

where $\mathbf{f}_i$, $\mathbf{g}_i$ are n-dimensional item-vectors, and $\mathbf{l}$ and $\mathbf{r}$ are n-dimensional position vectors, and L denotes list length or number of pairs stored in the memory vector $\mathbf{m}_{\mathrm{A}}$. Features values for all vectors are sampled from $N(0,\sigma^2)$, and then vectors are strictly normalized. To ensure that memory for the association co-varies with memory for its order, item-position, and item-item associations share an associative encoding strength α_i, which is a scalar value sampled from $N(\mu,\sigma_\alpha)$, and where σ_α, and μ are free parameters. As we elaborate in later sections, the model infers order by comparing a dot product between a correct item-position pair to the memory vector, and a dot product between an incorrect item-position pair and the memory vector. By increasing μ (which increases the average associative encoding strength), the model can increase the difference between these two dot products, which increases overall order discrimination.

Model Σ

Model Σ also represents position as two distinct vectors $\mathbf{l}$ and $\mathbf{r}$, but instead adds these positional vectors element-wise to their respective item vector,

$$\mathbf{m}_{\Sigma} = \sum_{i=1}^{L} \alpha_i ((\mathbf{f}_i + \mathbf{l}) * (\mathbf{g}_i + \mathbf{r})), \tag{3.2}$$

where L, α_i, $\mathbf{f}_i$, $\mathbf{g}_i$, $\mathbf{l}$, and $\mathbf{r}$ are identical to their definitions in equation 3.1, and $\mathbf{m}_\Sigma$ denotes the memory vector. After expanding this equation, $\mathbf{m}_\Sigma = \sum_{i=1}^{L} \alpha_i((\mathbf{f}_i * \mathbf{g}_i) + (\mathbf{g}_i * \mathbf{l}) + (\mathbf{f}_i * \mathbf{r}) + (\mathbf{l} * \mathbf{r}))$, we can see it is equivalent to model A (equation 3.1) with an additional noise term, $\mathbf{l} * \mathbf{r}$. This model also draws associative encoding strengths α_i from $N(\mu,\sigma_\alpha)$, and also assumes σ_α, and μ are free parameters. Thus, at a given value of σ_α, increased values of μ can increase order discrimination ability for this model with the same mechanisms as in model A.

Model ϕ

To encode order, each item gains a set of position features,

$$\mathbf{m}_\phi = \sum_{i=1}^{L} \alpha_i((\mathbf{f}_i \oplus \mathbf{l}) * (\mathbf{g}_i \oplus \mathbf{r})), \tag{3.3}$$

where $\mathbf{l}$ and $\mathbf{r}$ consist of n_p positional features that are concatenated (denoted by $\oplus$) onto item vectors $\mathbf{f}_i$ and $\mathbf{g}_i$ respectively, and $\mathbf{m}_\phi$ denotes the memory vector. Vectors $\mathbf{f}_i$ and $\mathbf{g}_i$ each consist of unique item features, and have $n - n_p$ dimensions to ensure that resulting dimensions of the full vector, with position features, is always equal to n. All feature values, including position features, are independently sampled from $N(0, \sigma^2)$, and item vectors, with position features, are strictly normalized. The number of positional features (n_p) is a free parameter. The model can infer order by comparing a dot product between a pair of items with correct position features to the memory trace, $(\mathbf{f}_i \oplus \mathbf{l}) * (\mathbf{g}_i \oplus \mathbf{r}) \cdot \mathbf{m}_\phi$, and a dot product between pair of items with incorrect position features to the memory trace, $(\mathbf{f}_i \oplus \mathbf{r}) * (\mathbf{g}_i \oplus \mathbf{l}) \cdot \mathbf{m}_\phi$. Parameter α_i is a scalar associative encoding strength sampled from $N(1, \sigma_\alpha)$, where σ_α is a free parameter, and L is defined as in previous models.

Model Π

Position/order is encoded as patterns of permutation,

$$\mathbf{m}_\Pi = \sum_{i=1}^{L} \alpha_i(p_l(\mathbf{f}_i) * p_r(\mathbf{g}_i)), \tag{3.4}$$

where $\mathbf{f}_i$ and $\mathbf{g}_i$ are n-dimensional item vectors, of which n elements are independently sampled from $N(0, \sigma^2)$. Vectors are then strictly normalized, and $\mathbf{m}_\Pi$ denotes the memory vector. A distinct pattern of permutation is applied to every left position item, denoted by p_l, and another pattern of permutation is applied to the right position item, denoted by p_r. The number of elements permuted by p_l and p_r is a free parameter n_{perm}. The model can infer order by comparing a dot product between a pair of items with the correct position permutations to the memory vector, $p_l(\mathbf{f}_i) * p_r(\mathbf{g}_i) \cdot \mathbf{m}_\Pi$, and a dot product between a pair of items with incorrect position permutations and the memory vector, $p_r(\mathbf{f}_i) * p_l(\mathbf{g}_i) \cdot \mathbf{m}_\Pi$. Parameters α_i, and L are defined as in equation 3.3.

3.2 Empirical benchmark 1: Order recognition and associative symmetry

We first wondered whether each of these models could produce above-chance order recognition performance, with a moderate relationship to cued recall performance, and alongside the near-perfect symmetry between forward and backward cued recall. Benchmark data was obtained from experiment 1 in Thomas, Ayuno, Kluger, and Caplan (in press),[4]. The design of this experiment is illustrated in Figure 3.2. Similar to Kato and Caplan (2017), participants first performed cued recall[5] of studied word pairs. This was followed by either order or associative recognition depending on condition. Departing from Kato and Caplan (2017), cued recall was only tested once per pair, meaning that the correlation between forward and backward cued recall could not be measured at the level of individual pairs, although we as we discuss below, we used general ranges from previous experiments to check whether the correlations in the present models are consistent with previous reports. Finally, Thomas et al. (in press) included an additional between-subject manipulation where some participants received instructions to use a memory strategy; however, we only used the control group in the following fits, as it was most comparable to conditions in Kato and Caplan (2017).

We derived three empirical benchmarks, 1a, 1b, and 1c from this data-set to evaluate models. Benchmark 1a was order recognition performance separated by cued recall correctness (and associative recognition as a control), alongside the near-perfect correlation between forward and backward cued recall. Benchmark 1b was individual differences in order recognition performance, that occupied a range around the means observed in benchmark 1a. Here we wondered if models could not only produce means that characterized empirical data, but account for individual participants within the data-set using different parameter sets. Benchmark 1c, was the *between-subject* correlation between order and cued recall. In both Kato and Caplan (2017) and Thomas et al. (in press), this correlation was well-below the correlation between associative recognition and cued recall. To distinguish benchmark 1c from benchmark 1a, we also denote benchmark 1a as the *within-subject* relationship between order recognition and cued recall.

To begin, we describe how order recognition, cued recall and associative recognition are simu-

[4] data is posted at `https://osf.io/x78gp/?view_only=17fbc3e1614545648d45ac19e62c2249`

[5] cued recall was performed in both the forward and backward direction, and direction was counterbalanced so that each participant received equal forward and backward cued recall trials

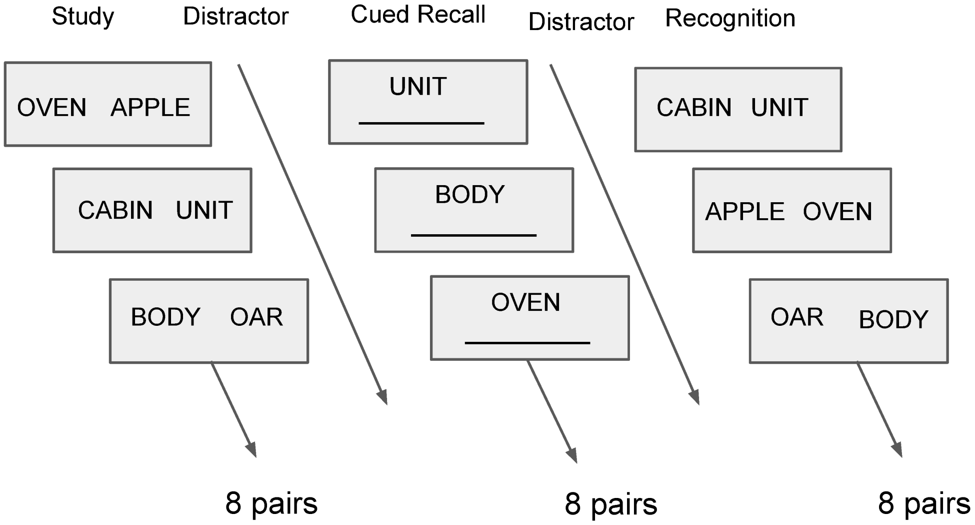


Figure 3.2: The design of one cycle from experiment 1 of Thomas et. al. (In Review). The recognition task performed (associative versus order) was a between-subject factor. This procedure was repeated for a total of eight cycles, after which participants completed other tasks and a questionnaire not pictured here.

lated in each of our models, examine the sensitivity of model performance to free parameters, and then begin evaluations against each benchmark in-sequence.

3.2.1 Simulation Methods

Encoding.

Each model encodes a memory vector **m** according to its respective encoding expression defined above (Equations 3.1–3.4). Each memory vector **m** consists of $L = 8$ unique pairings of 16 different items, matching Thomas et al. (in press).

Order recognition.

Two dot products are used to assess model order recognition performance. First, a dot product between items with the correct position and the memory trace, defined as follows for models A, Σ, ϕ, and Π respectively,

$$\iota_A = \quad ((\mathbf{f}_i * \mathbf{l}) + (\mathbf{g}_i * \mathbf{r})) \cdot \mathbf{m}_A, \tag{3.5}$$

$$\iota_\Sigma = \quad ((\mathbf{f}_i + \mathbf{l}) + (\mathbf{g}_i + \mathbf{r})) \cdot \mathbf{m}_\Sigma, \tag{3.6}$$

$$\iota_\phi = \quad ((\mathbf{f}_i \oplus \mathbf{l}) * (\mathbf{g}_i \oplus \mathbf{r})) \cdot \mathbf{m}_\phi, \tag{3.7}$$

$$\iota_\Pi = \quad (p_l(\mathbf{f}_i) * p_r(\mathbf{g}_i)) \cdot \mathbf{m}_\Pi, \tag{3.8}$$

And then, a dot product between items with incorrect positions and the memory trace, simulated for each of our four models using the following expressions respectively,

$$\rho_A = \quad ((\mathbf{f}_i * \mathbf{r}) + (\mathbf{g}_i * \mathbf{l})) \cdot \mathbf{m}_A, \tag{3.9}$$

$$\rho_\Sigma = \quad ((\mathbf{f}_i + \mathbf{r}) + (\mathbf{g}_i + \mathbf{l})) \cdot \mathbf{m}_\Sigma, \tag{3.10}$$

$$\rho_\phi = \quad ((\mathbf{f}_i \oplus \mathbf{r}) * (\mathbf{g}_i \oplus \mathbf{l})) \cdot \mathbf{m}_\phi, \tag{3.11}$$

$$\rho_\Pi = \quad (p_r(\mathbf{f}_i) * p_l(\mathbf{g}_i)) \cdot \mathbf{m}_\Pi, \tag{3.12}$$

For equations 3.7 and 3.11 (model ϕ), position features are first concatenated to item vectors, and then the entire vector is strictly normalized. For all other models (equations 3.5, 3.6, 3.8, 3.9, 3.10, and 3.12), all item vectors and position vectors are strictly normalized. Each equation above is computed for all L pairs in memory, which returns L samples per list. Overall order recognition sensitivity (d') is computed from these L samples across all lists according to $d' = \frac{(\mu_\iota - \mu_\rho)}{\sqrt{0.5(\sigma_\iota^2 + \sigma_\rho^2)}}$.

Cued recall

Cued recall is simulated with the correlation operation, denoted with #. Forward cued recall $\mathbf{f}_i \# \mathbf{m} \approx \mathbf{g}_i$, and backward cued recall, $\mathbf{g}_i \# \mathbf{m} \approx \mathbf{f}_i$, are simulated for all studied pairs. For each cued recall trial, a dot product is computed between the retrieved vector and all $L \times 2$ item vectors representing possible candidate responses, which we refer to as lexicon vectors. Lexicon vectors are strictly normalized. The highest match is selected as the response with a winner-take-all rule, and is scored correct if it matches the target item. Following Thomas et al. (in press), where cue words had no positional information, in all models position is excluded for the cue and lexicon vectors; 1) Model A, and Σ, positional vectors $\mathbf{l}$ and $\mathbf{r}$ are omitted from all cued recall operations. 2) Model ϕ, position features for cue vectors and lexicon candidate item vectors are replaced with noise, sampled from $N(0,\sigma^2)$ for each item. 3) Model Π, the cue vector and lexicon vectors are not permuted, departing from previous implementations (Kelly et al., 2013).

Associative recognition.

The following two dot products are used to assess model associative recognition performance,

$$\iota = \quad (\mathbf{f}_i * \mathbf{g}_i) \cdot \mathbf{m}, \tag{3.13}$$

$$\rho = \quad (\mathbf{f}_i * \mathbf{g}_x) \cdot \mathbf{m} \tag{3.14}$$

Equation 3.13 is a dot product between the memory vector and an old (studied) pairing of list items, and equation 3.14 is a dot product between the memory vector and a new pairing of list items. For equation 3.14, this dot product is repeated for L unique new pairings between left and right items from the studied list. All item vectors $\mathbf{f}_i$ and $\mathbf{g}_i$ are strictly normalized. Overall associative recognition performance (d') is computed using the outputs of these two dot products repeated for L pairs, across all lists, according to $d' = \frac{(\mu_\iota - \mu_\rho)}{\sqrt{0.5(\sigma_\iota^2+\sigma_\rho^2)}}$. Just as for cued recall, we assume no positional information in both equations; 1) Model A, and Σ, positional vectors $\mathbf{l}$ and $\mathbf{r}$ are omitted from intact and recombined probes. 2) For model ϕ, this means that position features for item vectors in intact and recombined probes are replaced with noise, sampled from $N(0,\sigma^2)$. 3) Model Π, item vectors in intact and recombined probes are not permuted.

Procedure

Following Thomas et al. (in press), encoding, cued recall, order recognition, and associative recognition are repeated for eight word lists. For recognition tasks, this results in $L\times 8$ intact probe matches (OR: Equations 3.5–3.8, AR: Equation 3.13), and reverse/recombined probe matches (OR: Equations 3.9–3.12, AR: Equation 3.14), from which order and associative recognition sensitivity (d') is computed. Similarly, cued recall accuracy was computed across $L\times 8$ trials for forward cued recall, and $L\times 8$ trials for backward cued recall.

3.2.2 Parametric plots of model performance.

Before fits to data, we wanted to understand the sensitivity of each model to parameters, with special attention to the single parameter that directly modifies each model's ability to discriminate order. This parameter was μ in models A and Σ, n_p in model ϕ, and n_{perm} in model Π. These parameters are now called the "order parameter" of each model. We simulated cued recall, order recognition, and associative recognition at the following values of each model's order parameter; for models A and Σ, $\mu = \{0, 0.1, 0.2 ...,1.0\}$, model ϕ, $\frac{n_p}{n} = \{0, 0.1, 0.2 ...,1.0\}$, and in model Π, $\frac{n_{perm}}{n} = \{0, 0.1, 0.2 ...,1.0\}$. Simulations were repeated for $\sigma_\alpha = \{$ 0.1, 0.5 ,1.0$\}$ (SD of associative encoding strength α). Total item vector features was held constant at $n = 100$ for all simulations, and all procedures were according to the specifications stated above.

Results

Parameter μ in models A and Σ had a positive relationship to performance in all memory tasks (figure 3.3). In contrast, parameter n_{perm} in model Π had a positive relationship with order recognition performance, but a negative relationship with associative recognition and cued recall performance. Parameter n_p in model ϕ was similar to n_{perm} in this way, but negatively affected order recognition performance after roughly half of the item vectors consisted of position features. With a few exceptions, reducing the value of σ_α, and therefore overall noise, improved performance for all models and memory tasks. Some of these relationships between model parameters and performance could be changed if models were implemented in different ways. For example, in model Π, we did not permute the cue vector based on position, because cue words did not contain position information in Thomas et al. (in press). However, if we did permute the cue vector, all of the features of the cue vector would be diagnostic for cued recall, rather than just the $n - n_p$ non-permuted features, and there would then be a positive relationship between n_{perm} and cued recall performance.

3.2.3 Empirical benchmark 1a: The moderate within-subject relationship between order recognition and cued recall correctness

Thomas et al. (in press) found that order recognition performance was significantly better when cued recall for that pair was correct, but well below associative recognition for correctly recalled word pairs (Figure 3.4). To test if each model could account for these within-subject patterns we performed quantitative fits to means in figure 3.4, along some other empirical constraints described below.

Given the challenge associative symmetry posed in previous efforts to modify models (see introduction), we included it as an additional constraint for the following model fits. We quan-

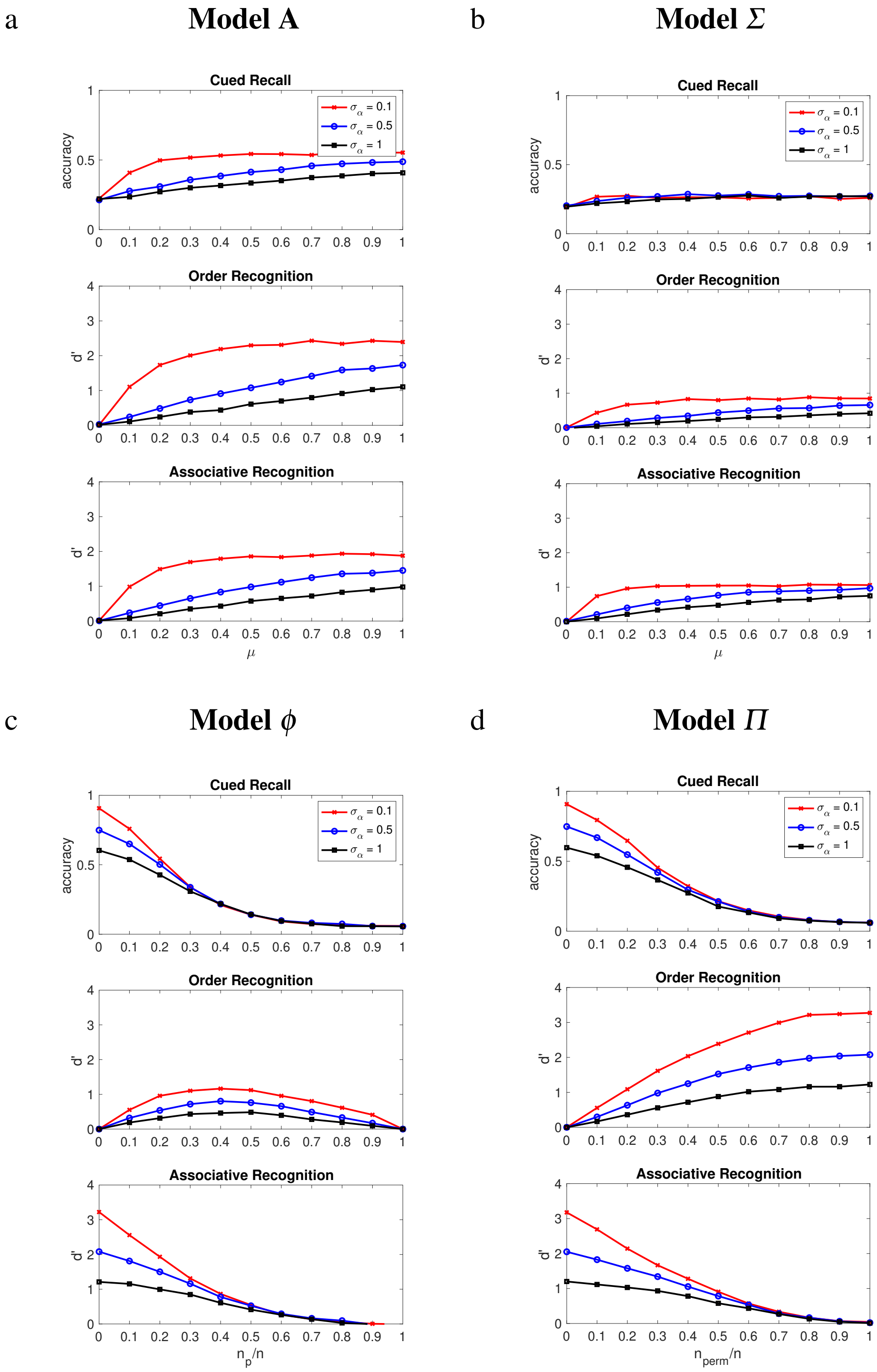


Figure 3.3: Parametric plots of cued recall, order recognition and associative recognition performance for each model as a function of; 1) models A and Σ: mean associative encoding strength μ, 2) model ϕ: number of position features (n_p), 3) model Π: number of permuted features (n_{perm}). Simulations were repeated for $\sigma_\alpha = \{\ 0.1, 0.5\ ,1.0\}$.

titatively fit the symmetry between forward and backward cued recall accuracy using data from Thomas et al. (in press). Previous studies (e.g., Kahana, 2002; Kato & Caplan, 2017) have used Yule's Q (Bishop, Fienberg, & Holland, 1975), to measure the within-pair correlation between forward and backward cued recall . Yule's Q quantifies the relationship between two tests with dichotomous outcomes, and like a Pearson correlation, ranges from -1 to 1. Thomas et al. (in press) could not compute Yule's Q because pairs were only tested with cued recall once. However, given that high Yule's Q more diagnostic of associative symmetry than average performance (Kahana, 2002), we still checked whether each model could produce values in the basic empirical range, such as Q $\approx$ .85 in Kato and Caplan (2017).

Parameter search and methods

Each model was fit to the following empirical values from Thomas et al. (in press); 1) order recognition d' for correctly recalled pairs, 2) order recognition d' for incorrectly recalled pairs, 3) associative recognition d' for correctly recalled pairs, 4) associative recognition d' for incorrectly recalled pairs, 5) the *difference* between order recognition d' for correct and incorrectly recalled pairs, 6) the *difference* between associative recognition d' for correct and incorrectly recalled pairs, 7) forward cued recall accuracy, 8) backward cued recall accuracy. The empirical values for each of these measures are reported in Table 3.2.

The closest fit for each model was determined via direct search. The direct search matrix was defined across the following parameters and parameter ranges; (1) $\sigma_\alpha = \{0, 0.1, 0.2 ...,1.0\}$, (2) $n = \{10, 20, 30 ...,500\}$. (3) Again, the third free parameter was specific to each model. For models A and Σ, μ (mean of associative encoding strength)= $\{0, 0.1, 0.2 ...,1.0\}$, model ϕ, number of positional features, $\frac{n_p}{n} = \{0, 0.1, 0.2 ...,1.0\}$, and in model Π, the number of permuted features, $\frac{n_{perm}}{n} = \{0, 0.1, 0.2 ...,1.0\}$. Forward cued recall, backward cued recall, order recognition and associative recognition were simulated as described above, for 8 lists of L pairs. These simulations were iterated 100 times for each cell of the direct search matrix, and model predicted values were averaged across these 100 iterations. Root-Mean-Squared Error (RMSE) was computed between empirical and model predicted values for the four means plotted in Figure 3.4. RMSE was then transformed to Bayesian Information Criterion (BIC) values via an estimation of log-likelihood (Burnham & Anderson, 2004). The BIC minimum was selected from the direct search matrix to find the best fitting parameter set. By convention, if ΔBIC > 2 the models are considered meaningfully different.

To compute model predictions for Yule's Q, we used the outcome of cued recall simulations for each pair in the backward and forward direction. Predicted Yule's Q values were generated for each of the 100 iterations at each cell of the direct search matrix as follows. The frequency of the following four outcomes was tallied; a = # of pairs where forward and backward cued recall were correct, b = # of pairs where forward cued recall was correct and backward cued recall was incorrect, c = # of pairs where forward cued recall was incorrect and backward cued recall was correct, d = # of pairs where both backward and forward cued recall were incorrect. Yule's Q is then calculated according to $(ad-bc)/(ad+bc)$, and can range from -1 to 1. Yule's Q values for each cell were then log-odds transformed, averaged, and then inverse log-odds-transformed[6] to generate a single predicted value at each cell of the direct search matrix.

For comparison we also plotted and reported the performance of a reference model by simulating 100 iterations of the Model Π at $\frac{n_{perm}}{n} = 0$, $\sigma_\alpha = 1$, and $n = 100$. At these parameter values this model is equivalent to an unmodified convolution model with no information for item position/order. This model is unable to produce order recognition d' above 0, and would thus be unconstrained by empirical order recognition performance. As a result, we did not fit the reference model to data.

Results

All four modified convolution models improved substantially on the reference model fits (Table 3.1), but Model A (item-position associations) produced worse quantitative fits than models Σ and Π, although these differences fell short of meaningful ($\Delta BIC > 2$). Nonetheless, all models, including model A, could produce greater order recognition performance for correctly recalled pairs, that was also well below associative recognition for correctly recalled pairs (Figure 3.4).

Models were also quite successful at preserving associative symmetry. All models exhibited equal forward and backward cued recall accuracy; however, models over-predicted the magnitude of accuracy values. The one exception was model A, which produced values that aligned more with empirical observations. There was a marked reduction in Yule's Q for all models compared to the reference model (Table 3.2), although this could be considered a strength of the present models,

[6] This followed analyses of empirical Yule's Q in Kato and Caplan (2017), who log-odds transformed Yule's Q to ensure that these measures met parametric assumptions.

Table 3.1: Best fitting model parameters for fits to benchmark 1a. Data was obtained from experiment 1 in Thomas et. al. (in review). All models produced substantially closer fits compared to the reference model (BIC>2), but model A performed substantially worse than models Σ, ϕ and Π. For model ϕ, free parameters were σ_α, n_p, and n. for model Π, free parameters were σ_α, n_{perm}, and n, and for models A and Σ, free parameters were μ, σ_α, and n. For models ϕ and Π, parameter μ was held constant.

Model	n	σ_α	Order parameter	BIC
Reference	400	0.5	N/A	5.88
Model A	50	0.3	$\mu = 0.9$	–13.65
Model Σ	440	0.2	$\mu = 0.8$	–15.39
Model ϕ	300	0.3	$\frac{n_p}{n} = 0.3$	–14.99
Model Π	200	0.4	$\frac{n_{perm}}{n} = 0.3$	–15.64

as this was more comparable to values that are observed in behavioural data (Yule's Q > 0.85 for all groups in Kato & Caplan, 2017). In sum, all modifications extended convolution to support moderate order recognition, without compromising associative symmetry.

3.2.4 Empirical benchmark 1b: Fits to individual differences in order recognition performance

Fits to aggregate data can be informative, but can lead to misleading conclusions if participants vary substantially, where some participants are better fit by one model and others, by a different model. Indeed, even though order recognition performance exhibits a moderate relationship to cued recall, individual participants in Thomas et al. (in press) occupied a range around these mean values (Figure 3.6). Thus, we also tested how each model could fit individual differences.

Parameter search

We fit models to individual participant values for; 1) order recognition d' for correctly recalled pairs, 2) order recognition d' for incorrectly recalled pairs, 3) log-odds transformed cued recall accuracy.[7] Re-using simulated model predictions from the direct search matrix for benchmark 1a, best fits were selected by minimizing BIC for each participant.

[7] Log-odds transformed cued recall was included in the fitness measure because we also used the following model fits for benchmark 1c (see below).

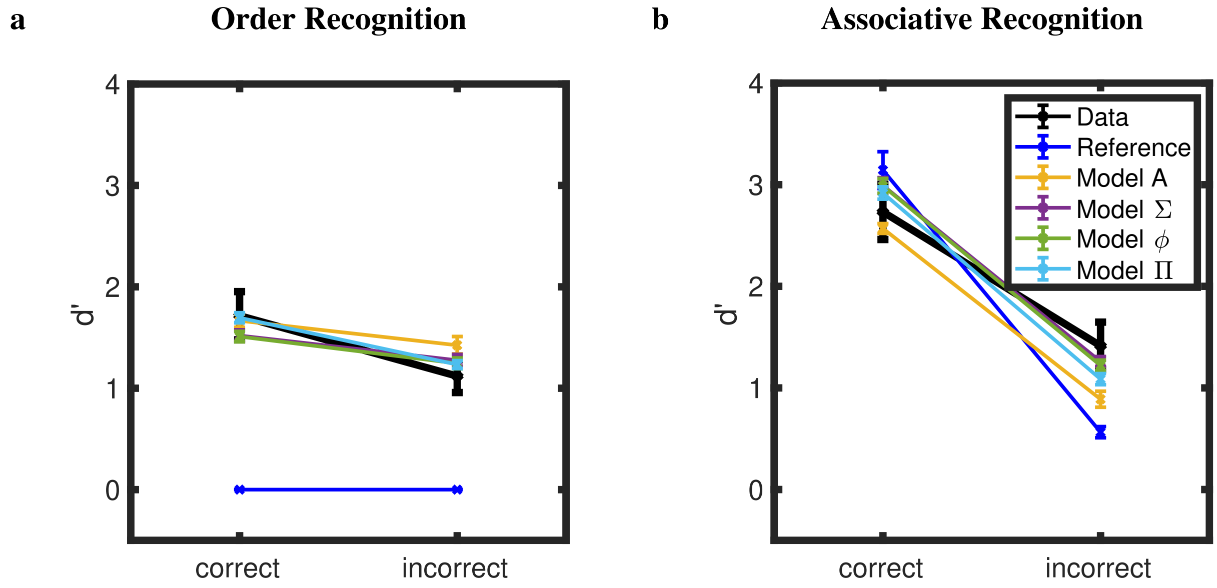


Figure 3.4: Each model's best fit to order and associative recognition d', for correct versus incorrectly recalled pairs, from Thomas et. al. (submitted). (a) Order recognition performance separated by correctness of recall. (b) Associative recognition performance separated by correctness of recall. Error bars in all panels represent 95% confidence intervals based on standard error of the mean.

Table 3.2: Data from experiment 1 in Thomas et. al. (in review) along with predictions generated by each model at best fitting parameters. “correct” denotes recognition performance for correctly recalled pairs. “incorrect” denotes recognition performance for incorrectly recalled pairs. “difference” is recognition performance for correctly recalled pairs minus performance for incorrectly recalled pairs, which provides a measure of the dependence of that recognition task on cued recall performance. Also reported are model predicted values for Yule’s Q, which unlike other measures, was not quantitatively fit-to.

	Order recognition d'			Associative recognition d'			Cued recall accuracy		
	correct	incorrect	difference	correct	incorrect	difference	forward	backward	Yule’s Q
Data	1.71	1.12	0.59	2.73	1.41	1.32	0.44	0.40	-
Reference	0	0	0	3.21	0.38	2.83	0.91	0.91	.99
Model A	1.67	1.42	0.25	2.57	0.89	1.68	0.34	0.35	.84
Model Σ	1.52	1.28	0.25	2.99	1.26	1.73	0.62	0.62	.87
Model ϕ	1.51	1.24	0.27	2.99	1.22	1.77	0.63	0.63	.87
Model Π	1.70	1.23	0.47	2.92	1.09	1.83	0.62	0.62	.90

Results

Starting with model Σ, this model could only produce a narrow range of performance values and was the poorest at accounting for individual differences across all models. Model A and ϕ produced a range of predictions (Figure 3.5), although model A biased towards predicting high order recognition d' for correctly recalled pairs (relative to the central tendency of the empirical data), while model ϕ tended to predict values closer to the center or even on the lower end of empirical order recognition d' for correctly recalled pairs. Model Π produced widest range of predictions, and seemed the most accurate at accounting for broadest range of individual differences.

As an additional step, we compared model fits with winner-take-all rule, tallying the number of times each model produced the lowest BIC value for a given participant. If a model did not win by significant margin compared to the other three models (ΔBIC > 2) we omitted that participant from plots and reported counts. 14 participants were excluded on this basis.

Model Π provided the strongest account of benchmark 1b, producing the best fit to 32 participants which were located throughout the scatter (Figure 3.6). This was followed by model ϕ which provided the best to 6 participants. Next was model A, which provided the best fit to 4

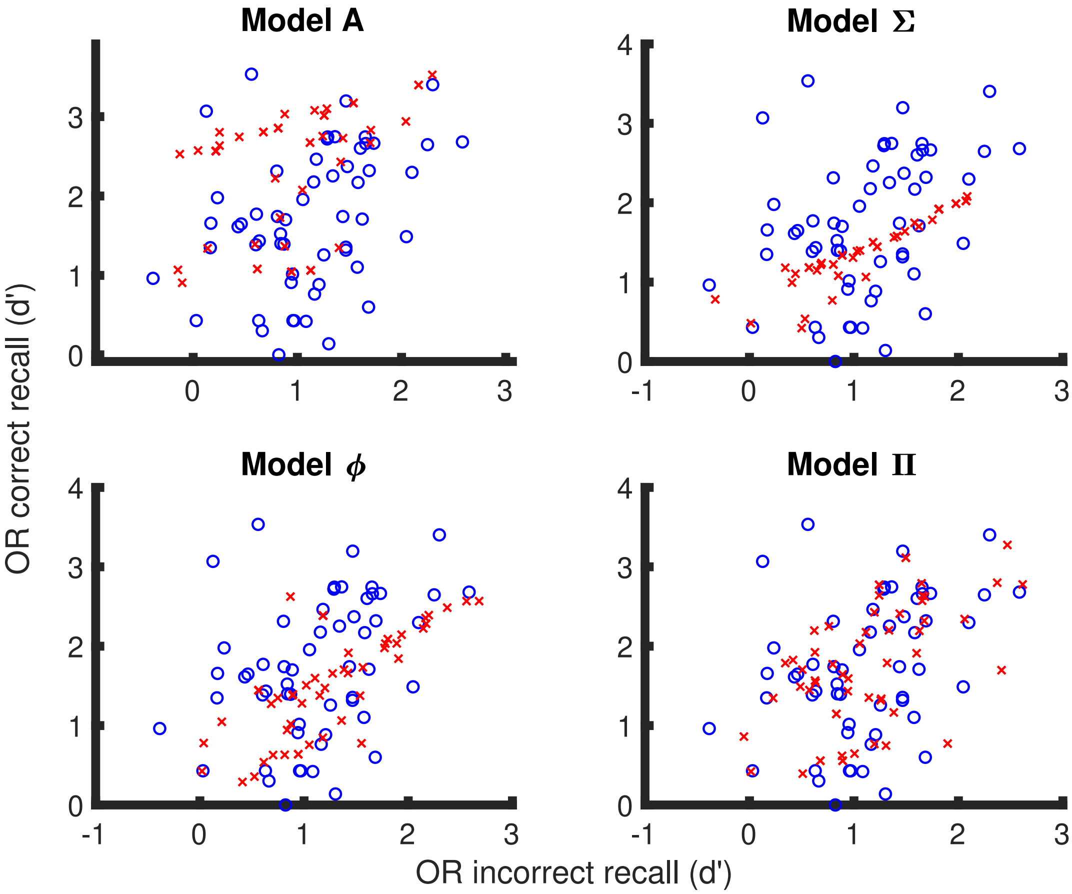


Figure 3.5: Model fits to individual participants values for order recognition d' for correct versus incorrectly recalled pairs from Thomas et. al. (in review).

participants that were located in the upper regions of the scatter plot (with high OR d' for correctly recalled pairs). Finally, model Σ did not win for any participant, strengthening the conclusion that this model provided a poor account of individual differences. In sum, in addition to providing good accounts of mean order recognition performance, model Π provided the closest fit to the largest number of individuals. However, certain participants were better described by other models, suggesting that participants may, in fact, judge order in more than one way.

Distribution of best fitting model parameters

Models used a wide range of parameter values to fit to individual participants, with roughly even distribution across the total explored parameter space (Figure 3.7, Table 3.3). There were some notable exceptions, for example, parameter distributions for n and σ_α in model Σ were noticeably

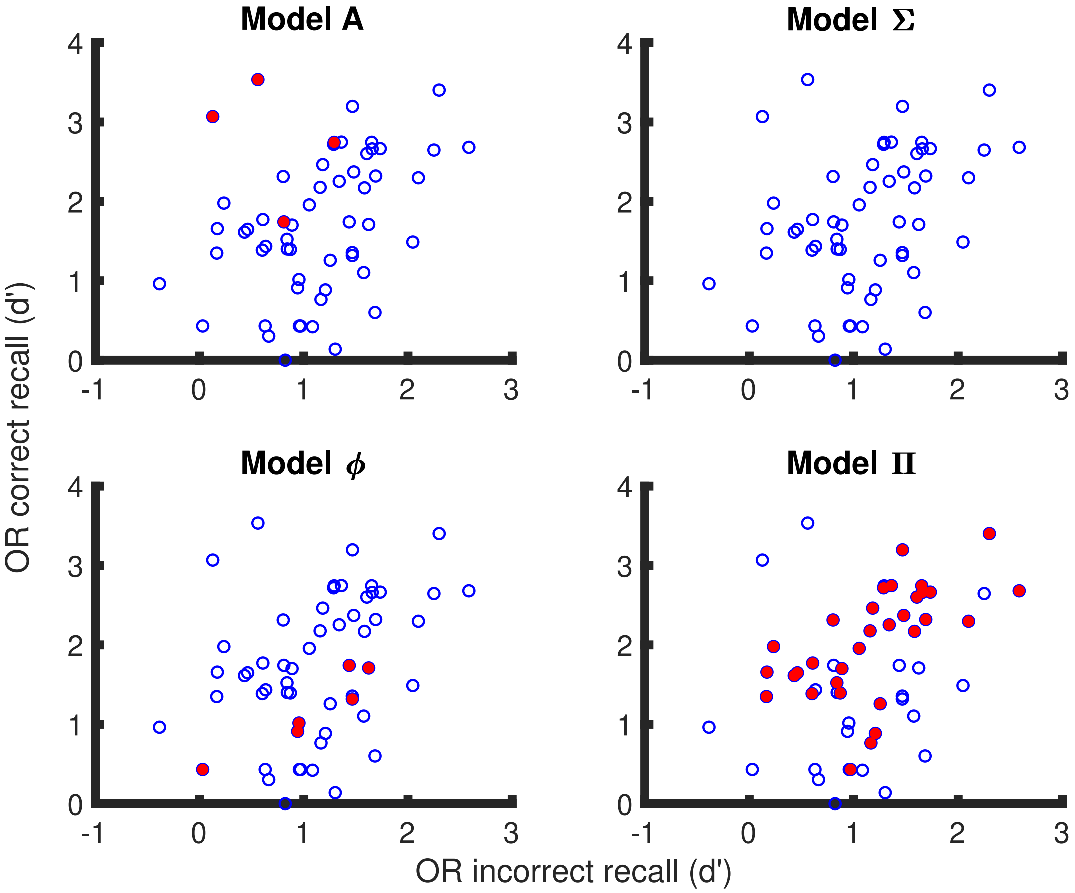


Figure 3.6: Model fits where filled-in circles denote the participants that each model provided the closest fit to by a margin of BIC> 2. Open circles also include the 30 participants that did not have clear winner.

Table 3.3: The mean and standard deviation (Mean(SD)) for the distribution of each model's free parameters used for fits to individual participants in benchmark 1b, along with Mean(SD) of the distribution of BIC values. This distribution of parameters were also applied to benchmark 1c.

Model	n	σ_α	Order parameter	BIC
Model A	294 (209)	0.43 (0.23)	$\mu = 0.69(0.26)$	–0.76 (5.46)
Model Σ	429 (141)	0.21 (0.19)	$\mu = 0.74(0.25)$	1.17 (4.29)
Model ϕ	379 (98)	0.17 (0.24)	$\frac{n_p}{n} = 0.56(0.28)$	–6.59 (5.11)
Model Π	329 (121)	0.57 (0.32)	$\frac{n_{perm}}{n} = 0.4(0.25)$	–10.66 (5.18)

skewed. These distributions indicate model Σ used high values of n and low values of σ_α to fit participants, both of which reduce the overall noise in the memory trace. Interestingly, model Σ had an additional noise term compared to model A, and may have used parameters n and σ_α to counteract this effect. For model ϕ, the distribution of parameter σ_α was also right skewed, although the other two free parameters in this model were more evenly distributed (Figure 3.7).

3.2.5 Empirical benchmark 1c: Between-subject correlations between recognition and recall extrapolated from fits to benchmark 1b

Thomas et al. (in press) also examined *between-subject* correlations between both recognition tasks and cued recall performance. These were consistent with benchmark 1a; there was a moderate correlation between order recognition and cued recall performance, but this was well below the correlation between associative recognition and cued recall (Figure 3.8). We wondered if models would also exhibit a moderate between-subject relationship between cued recall and order recognition. Rather than re-fit the models, we simply plotted model output from previous fits to benchmark 1b. Note, this meant that plotted model predictions for associative recognition d' were not included in the original fitness measure at all. This placed models at a significant disadvantage when producing the associative recognition-cued recall correlation, especially considering that a completely different set of participants were tested for associative recognition in Thomas et al. (in press).

Results

For models ϕ and Π the order recognition-cued recall correlation was smaller than the associative recognition-cued recall correlation, but fell short in accounting for the magnitude of correlations observed in behaviour (Figure 3.8). Models A and Σ produced order recognition-cued recall correlations that were comparable to the associative recognition-cued recall correlation, essentially

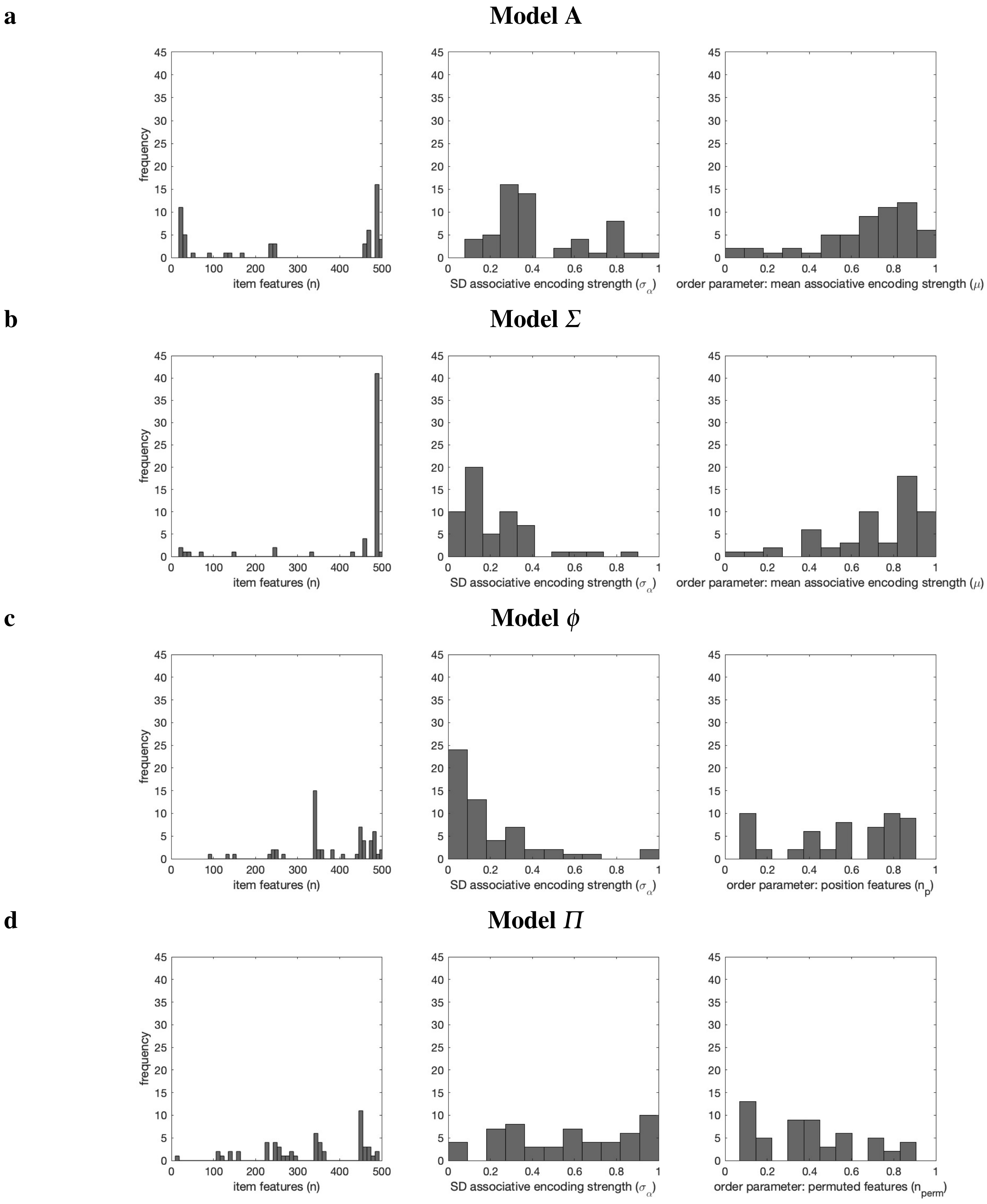


Figure 3.7: Histograms for the distributions of each model's free parameters for fits to benchmark 1b and 1c. Panels on the left plot the distribution of total item-features (n) used for model fits. Panels in the middle plot the distribution of SD associative encoding strength (σ_α) for model fits. Panels on the right plot the distribution of each model's order parameter values (μ,n_p,n_{perm}).

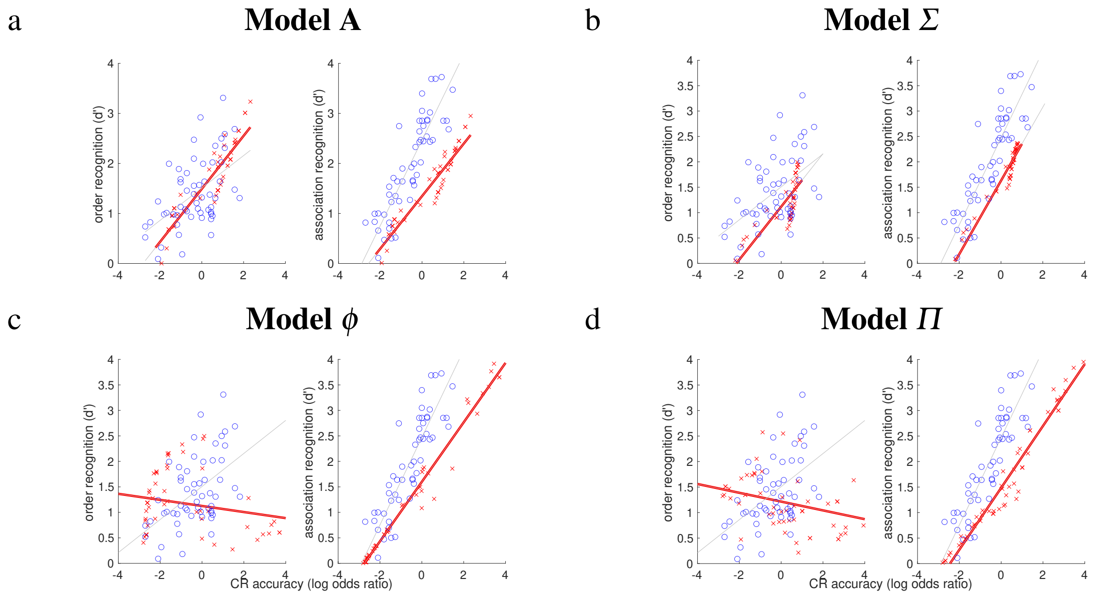


Figure 3.8: Model predictions and empirical data for order recognition versus log-odds transformed cued recall accuracy (left panels), and also associative recognition versus log-odds transformed cued recall accuracy (right panels). Model predicted values were generated in fits to benchmark 1b. Least squares lines for model predicted values are plotted in red, and least square lines for behavioural data are plotted in light grey. Each circle represents a participant, and crosses represents model predicted values.

predicting a maximal relationship between order recognition and cued recall. Despite the mismatch between model predictions and the magnitude of each correlation in behaviour, because models were not quantitatively fit to this data, the conclusion here is not that models are unable to account for benchmark 1c. In fact, although we do not report it here, models were quite good at producing the behavioural values for each correlation when directly fit to quantitative values for each participant. Rather, the conclusion here is that models cannot account for benchmark 1c by extrapolating from fits to benchmark 1b.

3.3 Empirical benchmark 2: double-function lists

Although models varied in their ability to account for individual differences, we have shown that simple modifications to convolution can produce moderate order memory without compromising its inherent symmetry. As a further test of each model, we leveraged another paradigm that demands memory for constituent-order, double function lists (AB..., BC..., CA...). Recall that standard convolution models are unable to disambiguate double function pairs during cued recall (see introduction). For example, if A is presented as a cue to a convolution model, both B and C are retrieved equally, and the model must guess. We start with algebraic expressions to come to general conclusions about how each model may solve this task, and then tested these conclusions with simulations.

Model A

First assume that three double-function pairs are encoded in memory, AB, BC, and CA. Following Rehani and Caplan (2011), each item appears in two pairs, exactly once in the left position and exactly once in the right position. This is expressed in model A as follows,

$$\begin{aligned}\mathbf{m} = \mathbf{a} * \mathbf{b} + \mathbf{a} * \mathbf{l} + \mathbf{b} * \mathbf{r}\\ + \mathbf{b} * \mathbf{c} + \mathbf{b} * \mathbf{l} + \mathbf{c} * \mathbf{r}\\ + \mathbf{c} * \mathbf{a} + \mathbf{c} * \mathbf{l} + \mathbf{a} * \mathbf{r}\end{aligned} \tag{3.15}$$

where **a**, **b**, and **c** denote item vectors, and **l** and **r** denote left and right position vectors respectively. Cued recall is expressed as follows,

$$\mathbf{a} \# \mathbf{m} = \mathbf{b} + \mathbf{c} + \mathbf{l} + \mathbf{r} \tag{3.16}$$

We see that the retrieved vector is essentially a sum of **b** and **c** with noise. As a result, there is no information to help the model select between competing items, resulting in perfect double function interference. To address this, one could incorporate the positional vector into the cue,

$$(\mathbf{a} + \mathbf{l}) \# \mathbf{m} = 2\mathbf{b} + 2\mathbf{c} + \mathbf{a} + \mathbf{l} + \mathbf{r} \tag{3.17}$$

However, the retrieved vector is still equally similar to **c** and **b**. This is because the positional vector **l** is associated to every item exactly once in the list, and provides no information to solve double function interference.

Model *Σ*

Model *Σ* cannot solve double function interference for the same reason. Again, assume that pairs AB, BC, and CA are encoded in memory,

$$\begin{aligned}\mathbf{m} = (\mathbf{a}+\mathbf{l}) * (\mathbf{b}+\mathbf{r}) \\ +(\mathbf{b}+\mathbf{l}) * (\mathbf{c}+\mathbf{r}) \\ +(\mathbf{c}+\mathbf{l}) * (\mathbf{a}+\mathbf{r})\end{aligned} \tag{3.18}$$

Expanding the above expression shows that model Σ is equivalent to model A, with an additional noise term $(\mathbf{l} * \mathbf{r})$ generated for each pair,

$$\begin{aligned}\mathbf{m} = (\mathbf{a}*\mathbf{b}) + (\mathbf{a}*\mathbf{r}) + (\mathbf{b}*\mathbf{l}) + (\mathbf{l}*\mathbf{r}) \\ +(\mathbf{b}*\mathbf{c}) + (\mathbf{b}*\mathbf{r}) + (\mathbf{c}*\mathbf{l}) + (\mathbf{l}*\mathbf{r}) \\ +(\mathbf{c}*\mathbf{a}) + (\mathbf{c}*\mathbf{r}) + (\mathbf{a}*\mathbf{l}) + (\mathbf{l}*\mathbf{r})\end{aligned} \tag{3.19}$$

In its expanded form, we can see that both positional vectors are associated to every item in the list. As a result, if cued recall is simulated with the cue $\mathbf{a}+\mathbf{l}$,

$$(\mathbf{a}+\mathbf{l}) \,\#\, \mathbf{m} = 2\mathbf{b} + 2\mathbf{c} + \mathbf{a} + \mathbf{l} + 4\mathbf{r} \tag{3.20}$$

the retrieved vector is equally similar to the target item **b**, and non-target item **c**. Just as in model A, positional vector **l** provides no diagnostic ability.

Model ϕ

Position features also cannot be used to solve double function interference. AB, BC, and CA would be encoded as follows,

$$\mathbf{m} = (\mathbf{a} \oplus \mathbf{l} * \mathbf{b} \oplus \mathbf{r}) + (\mathbf{b} \oplus \mathbf{l} * \mathbf{c} \oplus \mathbf{r}) + (\mathbf{c} \oplus \mathbf{l} * \mathbf{a} \oplus \mathbf{r}) \tag{3.21}$$

Assuming $n = 3$ and $n_p = 1$, this can also be expressed in its expanded form,

$$\begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix} = \left(\begin{bmatrix} a_1 \\ a_2 \\ l_3 \end{bmatrix} * \begin{bmatrix} b_1 \\ b_2 \\ r_3 \end{bmatrix} \right) + \left(\begin{bmatrix} b_1 \\ b_2 \\ l_3 \end{bmatrix} * \begin{bmatrix} c_1 \\ c_2 \\ r_3 \end{bmatrix} \right) + \left(\begin{bmatrix} c_1 \\ c_2 \\ l_3 \end{bmatrix} * \begin{bmatrix} a_1 \\ a_2 \\ r_3 \end{bmatrix} \right) \tag{3.22}$$

$$\begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix} = \begin{bmatrix} a_1b_1 + a_2r_3 + l_3b_2 + & b_1c_1 + b_2r_3 + l_3c_2 + & c_1a_1 + c_2r_3 + l_3a_2 \\ a_1b_2 + a_2b_1 + l_3r_3 + & b_1c_2 + b_2c_1 + l_3r_3 + & c_1a_2 + c_2a_1 + l_3r_3 \\ a_1r_3 + a_2b_2 + l_3b_1 + & b_1r_3 + b_2c_2 + l_3c_1 + & c_1r_3 + c_2a_2 + l_3a_1 \end{bmatrix} \tag{3.23}$$

We can see that positional features l_3 and r_3 are distributed throughout the memory vector after convolution, appearing in terms with item features from every item of the list. Ultimately, this means that positional features are no longer specific to any item. This becomes clearer if we proceed with cued recall, which is expressed as, $\mathbf{x} = (\mathbf{a} \oplus \mathbf{l}) \# \mathbf{m}$, where $\mathbf{x}$ is the retrieved vector. The expanded form of equation is expressed as follows,

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ l_3 \end{bmatrix} \# \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix} \tag{3.24}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} a_1 m_1 + a_2 m_2 + l_3 m_3 \\ l_3 m_1 + a_1 m_2 + a_2 m_3 \\ a_2 m_1 + l_3 m_2 + a_1 m_3 \end{bmatrix} \tag{3.25}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1(a_1^2 + a_2^2 + l_3^2) + c_1(a_1^2 + a_2^2 + l_3^2) + a_1 l_3^2 + noise \\ b_2(a_1^2 + a_2^2 + l_3^2) + c_2(a_1^2 + a_2^2 + l_3^2) + a_2 l_3^2 + noise \\ r_3(a_1^2 + a_2^2 + l_3^2) + r_3(a_1^2 + a_2^2 + l_3^2) + r_3 l_3^2 + noise \end{bmatrix} \tag{3.26}$$

The retrieved vector $\mathbf{x}$ is essentially an equal sum of $\mathbf{b} \oplus \mathbf{r}$ and $\mathbf{c} \oplus \mathbf{r}$, and to a lesser extent $\mathbf{c} \oplus \mathbf{r}$. As a result, dot products to both candidate items will be equal ($\mathrm{E}[\mathbf{x} \cdot (\mathbf{b} \oplus \mathbf{r})] = \mathrm{E}[\mathbf{x} \cdot (\mathbf{c} \oplus \mathbf{r})]$), regardless of the number of positional features n_p. Thus, because position features are repeated for multiple items, they cannot be used to cue a specific item in memory. Thus, the position-feature model cannot solve double function interference.

Model *Π*

Permutation, in contrast, can be used solve double function interference. First assume that pairs AB, BC, and CA are encoded as follows,

$$\mathbf{m} = p_l(\mathbf{a}) * p_r(\mathbf{b}) + p_l(\mathbf{b}) * p_r(\mathbf{c}) + p_l(\mathbf{c}) * p_r(\mathbf{a}) \tag{3.27}$$

To understand why interfering pairs can be disambiguated with permutation, consider a case where the whole item is permuted ($\frac{n_{perm}}{n} = 1$) before encoding. Given this, $p_r(\mathbf{a})$ and $p_l(\mathbf{a})$ will behave as distinct, orthogonal items (assuming large n). As a result, if cued recall proceeds with the following expression,

$$p_l(\mathbf{a}) \# \mathbf{m} \tag{3.28}$$

$p_l(\mathbf{a})$ will only evoke pair, $p_l(\mathbf{a}) * p_r(\mathbf{b})$in memory, and grant the model perfect ability to disambiguate double function pairs.

If we assume only a subset of item vectors are permuted ($\frac{n_{perm}}{n} < 1$), The degree to which vector $p_l(\mathbf{a})$ retrieves the target $p_r(\mathbf{b})$ is proportional to n_{perm}. This is because the non-permuted portion of $p_l(\mathbf{a})$ is identical to the non-permuted portion of $p_r(\mathbf{a})$, it will also evoke pair $p_l(\mathbf{c}) * p_r(\mathbf{a})$. As we demonstrate below, changing n_{perm} allows the position-specific permutation model to produce a range of performance values ranging from zero (like an unmodified convolution model) to perfect (like a matrix model) ability to solve double function pairs.

3.3.1 Simulation methods

To test the insights gained from algebraic expressions, we also simulated double function lists with each of our four models.

Encoding

Assume that each model stores double function pairs AB, BC, and CA. Encoding for each model proceeds as follows,

$$\mathbf{m}_\phi = \alpha_1(\mathbf{a} \oplus \mathbf{l} * \mathbf{b} \oplus \mathbf{r}) + \alpha_2(\mathbf{b} \oplus \mathbf{l} * \mathbf{c} \oplus \mathbf{r}) + \alpha_3(\mathbf{c} \oplus \mathbf{l} * \mathbf{a} \oplus \mathbf{r}) \tag{3.29}$$

$$\mathbf{m}_\Pi = \alpha_1(p_l(\mathbf{a}) * p_r(\mathbf{b})) + \alpha_2(p_l(\mathbf{b}) * p_r(\mathbf{c})) + \alpha_3(p_l(\mathbf{c}) * p_r(\mathbf{a})) \tag{3.30}$$

$$\mathbf{m}_A = \alpha_1(\mathbf{a} * \mathbf{b} + \mathbf{a} * \mathbf{l} + \mathbf{b} * \mathbf{r}) + \alpha_2(\mathbf{b} * \mathbf{c} + \mathbf{b} * \mathbf{l} + \mathbf{c} * \mathbf{r}) + \alpha_3(\mathbf{c} * \mathbf{a} + \mathbf{c} * \mathbf{l} + \mathbf{a} * \mathbf{r}) \tag{3.31}$$

$$\mathbf{m}_\Sigma = \alpha_1((\mathbf{a} + \mathbf{l}) * (\mathbf{b} + \mathbf{r})) + \alpha_2((\mathbf{b} + \mathbf{l}) * (\mathbf{c} + \mathbf{r})) + \alpha_3((\mathbf{c} + \mathbf{l}) * (\mathbf{a} + \mathbf{r})) \tag{3.32}$$

where **a**, **b**, and **c** are word vectors with n features, **l**, **r** represent positional vectors in models A and Σ with n features, and α_1, α_2, and α_3 represent associative encoding strengths.

Cued recall

Assuming that A is a left-position cue, cued recall proceeds as follows,

$$(\mathbf{a} \oplus \mathbf{l}) \# \mathbf{m}_\phi \tag{3.33}$$

$$p_l(\mathbf{a}) \# \mathbf{m}_\Pi \tag{3.34}$$

$$(\mathbf{a} + \mathbf{l}) \# \mathbf{m}_A \tag{3.35}$$

$$(\mathbf{a} + \mathbf{l}) \# \mathbf{m}_\Sigma \tag{3.36}$$

Then for all models, a dot product is computed between the retrieved vector, and each of the vectors, **b**, and **c**, which represent candidate items B and C. Note that for model $\mathit{\Pi}$, the output of equation 3.34 is permuted with the inverse of the right permutation pattern to reproduce the original non-permuted item, following previous implementations of permutation (Jones & Mewhort, 2007; Kelly et al., 2013). If a model can disambiguate double function pairs, the retrieved item should be more similar to item B than to item C.

Procedure

For item vectors, **a**, **b**, and **c**, and position vectors **l**, and **r**, $n = 100$. Vector features were drawn from $N(0, \sigma^2)$, where $\sigma^2 = \frac{1}{n}$. Associative encoding strengths (α_1,α_2,α_3) were drawn from $N(\mu, \sigma_\alpha)$, where $\sigma_\alpha = 1$, and $\mu = 1$ for models ϕ and $\mathit{\Pi}$. We varied the number of item position features (n_p) in model ϕ, permuted features (n_{perm}) in model $\mathit{\Pi}$, and mean associative encoding strength (μ) in models A and $\mathit{\Sigma}$ according to the following ranges $\frac{n_p}{n} = \{0, 0.1, 0.2 ...,1.0\}$, $\frac{n_{perm}}{n} = \{0, 0.1, 0.2 ...,1.0\}$, $\mu = \{0, 0.1, 0.2 ...,1.0\}$. For each model, dot products between the retrieved vectors from equations 3.33-3.36, and candidate items **b** and **c** were averaged across 10000 iterations, for each value of n_p, n_{perm}, and μ.

3.3.2 Results

The main results from these simulations are plotted in figure 3.9. Confirming our arguments above, models ϕ, A, and $\mathit{\Sigma}$, could not solve interference between **b** and **c**, even when parameters $\frac{n_p}{n}$ and μ were increased. In contrast, for model $\mathit{\Pi}$ the difference in matching strengths between the retrieved vector to both **b** and **c** increased with parameter $\frac{n_{perm}}{n}$. At $\frac{n_{perm}}{n} = 1$, this difference reached the maximum possible value, where the matching strength to **c** reached the minimum dot product between two normalized vectors (≈ 0). This indicates that model $\mathit{\Pi}$ is able to mimic both zero, and perfect ability to solve double function interference, and all values in between. Taken together, this confirms the idea that models ϕ, A, and $\mathit{\Sigma}$ suffer from cue-overload when tested with cued recall for stored double function pairs. Permutation (model $\mathit{\Pi}$) overcame this challenge because permuting a given item by two different patterns (e.g., $p_l(\mathbf{a})$ versus $p_r(\mathbf{a})$) decreases similarity between both versions, in proportion to the amount of permuted features. This meant that a cue vector with a certain positional permutation can selectively activate a pair without activating the corresponding double function pair.

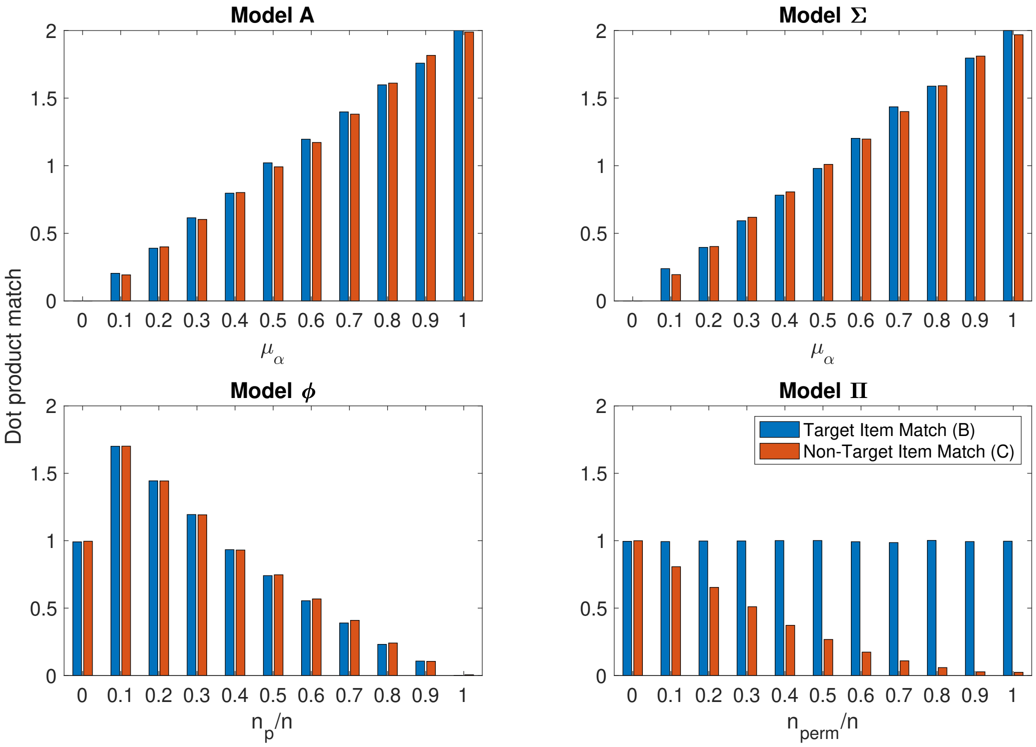


Figure 3.9: Double function list simulations for each model. Dot products were computed between a retrieved vector from cued recall, and the target versus the non-target item. For models ϕ, A, and Σ matching strengths are identical for the target and non-target at all parameter values. For model Π, the difference between the target and non-target item match, and therefore the ability to solve interference, increases with the proportion of permuted features. At $n_{perm}/n = 0$, model Π is equivalent to an unmodified convolution model and has no ability to solve this interference. At $n_{perm}/n = 1$, model Π is essentially non-commutative like in previous implementations of permutation (e.g., Kelly et. al., 2013), and has perfect ability to solve interference.

Relating these simulations back to previous model fits, when we fit models to averaged order recognition data (benchmark 1a), model Π achieved its best fit at $\frac{n_{perm}}{n} = 0.3$. At this same parameter value in the present simulations, model Π shows a clear separation between the target item and non-target item, but the match to the non-target item is not zero, consistent with high rates of errors observed to the non-target item in behaviour (Rehani & Caplan, 2011). In other words, model Π may not need to deviate from the fits to the other benchmark data to be able to perform well on double function lists.

3.4 Discussion

We started with the following puzzle: the perfect symmetry of convolution-based models matched behavioural data well, but offered no ability to discriminate the constituent-order of associations. This contradicted empirical data, which revealed that order recognition could be judged above-chance, and that this ability was moderately dependent on remembering the pairing itself.

All of our four models were constructed to address these challenges, and were largely successful. All models produced order recognition that was above chance, and moderately dependent on cued recall (Figures 3.3, 3.4, and 3.8). Model $\mathit{\Pi}$ provided the closest fit to the most amount of individual participant values for order recognition for correct versus incorrectly recalled pairs, although all models were capable of producing a range of performance values (benchmark 1b, Figure 3.6). When we extended fits from benchmark 1b to produce predicted values for between-subject correlations, models ϕ and $\mathit{\Pi}$ were able to produce smaller order recognition-cued recall correlations than associative recognition-cued recall correlations. Models did not match the magnitude of correlations seen in behaviour, but as we indicated above, models were not fit to order and associative recognition values. Finally, only model $\mathit{\Pi}$ could perform position-sensitive cuing without simultaneously retrieving every item in a position (Figure 3.9). We discuss the implications of these findings below.

Simple modifications can produce memory for order with symmetric associations We tested several possible mechanisms that could extend a convolution model to store order. Position permutation (model $\mathit{\Pi}$) produced close fits to order and associative recognition data at only 30% permuted features, departing from previous implementations (e.g. Jones & Mewhort, 2007; Kelly et al., 2013). The positional-feature model (model ϕ), item-position associations (model A) and addition of item and position vectors (model $\mathit{\Sigma}$) were similarly successful at fitting recognition data. This suggests convolution can be modified quite easily to produce moderate order recognition performance.

We also checked whether each model could preserve the inherent symmetry of convolution while fitting recognition data. This was especially important consider given the difficulty this additional constraint posed in previous efforts to modify matrix models. Matrix models start out asymmetric, but can be modified to produce associative symmetry by storing both the forward and backward associations (and with highly correlated forward and backward associative encoding strengths), although this removes any information for order. To regain some order, one could increase the forward association strength, but this causes the model to violate associative symmetry, along with generating additional erroneous predictions (see introduction). In contrast, all four of our models here maintained symmetry between forward and backward cued recall accuracy. Additionally, our models produced high Yule's Q values, although not quite as high as values from unmodified convolution model (Table 3.2).

Interestingly, less-than-perfect forward-backward Yule's Q may be more consistent with empirical findings. In data, the test/re-test correlation (both tests forward or both backward) is typically nearly perfect, whereas the forward-backward correlation is typically well below 1, around 0.8–0.9 (Kahana, 2002; Kato & Caplan, 2017; Rehani & Caplan, 2011; Rizzuto & Kahana, 2000, 2001; Sommer et al., 2008). In a symmetric model, one way to reduce the forward-backward correlation would be add noise between successive tests; however, this would also reduce the test/re-test correlation. In contrast, all four of our models produced correlations well below 1 without such a mechanism. Thus, it seems that deviating from the perfectly commutative convolution operation can also explain why forward and backward cued recall are slightly decoupled from one another (compared to testing twice in the same direction), without losing other desirable characteristics of convolution, such as equal accuracy in the forward and backward direction on average.

The success of our models shows that symmetric item-item associations can still support order judgements. Furthermore, the paradox between associative symmetry and moderate order memory may be particular to unmodified matrix model, which assume order is derived directly from a perfectly directional association. However, these results do not necessarily argue against matrix models, but suggest modifications to these models need to take a different approach. For example, one could incorporate partial-permutation into a symmetric matrix model as follows, $M = \alpha(p_l(\mathbf{a})p_r(\mathbf{b})^\intercal + p_r(\mathbf{b})p_l(\mathbf{a})^\intercal)$, where the forward and backward association share the same associative encoding strength α to produce high Yule's Q. The model could infer order by retrieving an item, then computing a dot product to a copy of this item with the correct position, $p_r(\mathbf{b}) \cdot (Mp_l(\mathbf{a}))$, and incorrect position, $p_l(\mathbf{b}) \cdot (Mp_l(\mathbf{a}))$. Order recognition performance, as the difference between these two dot products, would be proportional to amount of permuted features. This version of the matrix model may be able to function similarly to its cousin implemented with convolution (model $\mathit{\Pi}$).

Like convolution, some recent models within the REM framework such as Criss and Shiffrin

(2005) and Cox and Criss (2017, 2020) also disregard the order within associations. The ideas from models Π and ϕ could also be applied to these models quite easily. Indeed, Cox and Criss (2020) suggested something to this effect, where features representing the spatial locations of each item could be incorporated into item vectors in their model to produce some memory for order. Partial permutation could be applied to REM-based models with the same logic as with matrix models (described above). Because item-item associations are represented with concatenation within the REM framework, applying model A would be formally equivalent to applying model ϕ.

The influence of order/position on associative recognition Across six experiments and under various conditions, J. Yang et al. (2013) found that associative recognition probes were judged faster, and with higher accuracy when presented in the correct order, replicating the results of a number of studies (Giovanello, Schnyer, & Verfaellie, 2009; Haskins, Yonelinas, Quamme, & Ranganath, 2008; Wiegand, Bader, & Mecklinger, 2010). Our models here may help us understand how these results are still be consistent with symmetric associations in memory.

First, consider the position-specific permutation model (model Π). Assume the model stores the following pairs in memory, $\mathbf{m} = p_l(\mathbf{a}) * p_r(\mathbf{b}) + p_l(\mathbf{c}) * p_r(\mathbf{d})$. If the model "knows" that probes may be reversed at test, it is reasonable to assume that it will apply permutation to probe items to incorporate order into the recognition process. The model could simulate a "forward" intact trial as follows, $p_l(\mathbf{a}) * p_r(\mathbf{b}) \cdot \mathbf{m}$, along with a "backward" intact trial, $p_r(\mathbf{a}) * p_l(\mathbf{b}) \cdot \mathbf{m}$. These two matches are identical to our implementation of order recognition in model Π (equations 3.8 and 3.12), so we already know that the model can produce an advantage for forward intact probes. The model could not produce a forward advantage for recombined probes, because both $\mathrm{E}\left[p_l(\mathbf{a}) * p_r(\mathbf{d}) \cdot \mathbf{m}\right]$ and $\mathrm{E}\left[p_r(\mathbf{a}) * p_l(\mathbf{d}) \cdot \mathbf{m}\right]$ are equal to 0. Thus, the permutation model would predict that forward asymmetries for associative recognition are only driven by asymmetries in intact probe trials.

Model ϕ (position-features) would function similarly. Again, we know the model can produce a correct-order advantage for intact pairs because the comparison between a forward and backward intact trial is identical to its implementation of order recognition (Equation 3.7 and 3.11), thus $\mathrm{E}\left[(\mathbf{a} \oplus \mathbf{l}) * (\mathbf{b} \oplus \mathbf{r}) \cdot \mathbf{m}\right] > \mathrm{E}\left[(\mathbf{a} \oplus \mathbf{r}) * (\mathbf{b} \oplus \mathbf{l}) \cdot \mathbf{m}\right]$. However, because $\mathrm{E}\left[(\mathbf{a} \oplus \mathbf{l}) * (\mathbf{d} \oplus \mathbf{r}) \cdot \mathbf{m}\right]$ and $\mathrm{E}\left[(\mathbf{a} \oplus \mathbf{r}) * (\mathbf{d} \oplus \mathbf{l}) \cdot \mathbf{m}\right]$ are both equal to 0, like in model Π, this model predicts no order-advantage for recombined pairs.

Model A (the item-position association model), and by extension model Σ, can also produce a order-advantage for associative recognition. Given that encoding is as follows, $\mathbf{m} = (\mathbf{a}*\mathbf{b}+\mathbf{a}*\mathbf{l}+\mathbf{b}*\mathbf{r})+(\mathbf{c}*\mathbf{d}+\mathbf{c}*\mathbf{l}+\mathbf{d}*\mathbf{r})$, the model produces a correct-order advantage to intact probes because, $\mathrm{E}\left[(\mathbf{a}*\mathbf{b}+\mathbf{a}*\mathbf{l}+\mathbf{b}*\mathbf{r})\cdot\mathbf{m}\right] > \mathrm{E}\left[(\mathbf{a}*\mathbf{b}+\mathbf{a}*\mathbf{r}+\mathbf{b}*\mathbf{l})\cdot\mathbf{m}\right]$. However, unlike model Π and ϕ, model A (and Σ) can also produce a correct-order advantage for recombined probes, $\mathrm{E}\left[(\mathbf{a}*\mathbf{d}+\mathbf{a}*\mathbf{l}+\mathbf{d}*\mathbf{r})\cdot\mathbf{m}\right] > \mathrm{E}\left[(\mathbf{a}*\mathbf{d}+\mathbf{a}*\mathbf{r}+\mathbf{d}*\mathbf{l})\cdot\mathbf{m}\right]$.

Thus, models differ in their ability to produce correct-order advantages for recombined probes. In experiment 6 of J. Yang et al. (2013), associative recognition judgements were more accurate for intact pairs in the correct order (compared to incorrect order), but with no significant difference for recombined pairs. This may suggest that order did not influence judgements of recombined pairs, consistent with models Π and ϕ; however, because these were analyses of *accuracy* values, and not of d', they did not account for bias effects. In any case, our main point here is that symmetric associations can still cause associative recognition to depend on constituent-order if position/order is incorporated at encoding and at test.

Order incorporated into the item representation Models were largely comparable when fitting order recognition data. However, the double function task posed significant challenges to model A, model Σ, and model ϕ. This was because items in double function pairs appear in both positions exactly once. As a result, information for position is not specific to a given item, and incorporating position into cue vectors provided no information to select between interfering double function pairs. Model Π (partial permutations) could overcome this issue, because permuting a given item by two different patterns (e.g., $p_l(\mathbf{a})$ versus $p_r(\mathbf{a})$) decreases similarity between both versions, in proportion to the amount of permuted features, but permuting distinct items by the same pattern does not increase their similarity. This meant that a cue permuted based on position could activate a specific pair based on position, and the model solve double function interference.

The success of permutation here may allow us to come to more specific conclusions about how order is represented in memory, suggesting that order is encoded by directly modifying item representations based on position. The idea that item vectors are modified by their appearance in a word pair has precedence in existing memory models (Benjamin, 2010; Caplan et al., 2021; Criss & Shiffrin, 2004b; Cox & Criss, 2020). For example, Benjamin's (2010) DRYAD model encoded context as a subset of each item's features to explain age-related memory deficits in context memory. If we assume that order/position is also part of context, this idea would be quite similar to our position-feature model, although the present analysis shows the challenges with this representation choice when context is repeated for multiple items.

Caplan et al. (2021) proposed a model where certain features of a word were selectively attended to, while others were not, and set to zero. Attended features were based on the item it was paired with at study, implementing the idea certain meanings of a word are highlighted based on the context it appears in (e.g., BANK in RIVER BANK versus MONEY BANK). The model could use the pattern of attended features to judge pairings between items without storing any explicit associations. Vector permutation in our present model may be functionally related to this idea. Like permutation, setting certain features of an item to zero rotates item vectors in vector space, causing them to be dissimilar from the original word. Taken together with the success of permutation in the present article, this may suggest that a generality to permutation-like mechanisms to encode various information about the context in which that item appeared in, including its spatial position.

The influence of task demands and memory strategies on order-encoding strategies Although we found the most evidence for permutation with the present data, it is possible that other models could be supported under different conditions. For example, if participants studied word pairs, but were only required to judge their constituent-order (rather than item-item pairings), it might be optimal to ignore associations between items and focus on the relative positions of each item. In this case, a participant's cognitive strategy might be more consistent with the item-position association model (model A), or addition of item and position vectors (model Σ). Even in Thomas et al. (in press), model A provided a substantially closer fit ($BIC > 2$) to 4 participants (Figure 3.6), suggesting that even when association memory is tested, participants still may adopt qualitatively different order-encoding strategies that may consistent with different models. Future work could examine the conditions that cause participants to either encode order within the item representation, or as explicit item-position associations.

Non-commutative convolution in the brain Both Plate (2000) and Kelly, Mewhort, and West (2017) noted that precisely implementing convolution in the brain would require intricate patterns of neural connectivity. However, even if a network of neurons is wired perfectly to compute convolution, it is unlikely that the synaptic strengths would be perfectly equal within the network. Unequal synaptic strengths may actually be useful from the perspective of encoding order. Consider the following expression, $\mathbf{a} * \mathbf{b}$, which can be expanded as follows,

$$\begin{bmatrix} a_1b_1 + a_2b_3 + a_3b_2 \\ a_1b_2 + a_2b_1 + a_3b_3 \\ a_1b_3 + a_2b_2 + a_3b_1 \end{bmatrix} \tag{3.37}$$

Now consider a network of neurons that computes this operation, but by random chance, has one synapse that is stronger than the rest, represented with the coefficient ζ ,

$$\begin{bmatrix} m_{f1} \\ m_{f2} \\ m_{f3} \end{bmatrix} = \begin{bmatrix} a_1b_1 + \underline{\zeta a_2b_3} + a_3b_2 \\ a_1b_2 + a_2b_1 + a_3b_3 \\ a_1b_3 + a_2b_2 + a_3b_1 \end{bmatrix} \tag{3.38}$$

If same network computes the reversed association, $\mathbf{m}_b = \mathbf{b} * \mathbf{a}$,

$$\begin{bmatrix} m_{b1} \\ m_{b2} \\ m_{b3} \end{bmatrix} = \begin{bmatrix} b_1a_1 + \underline{\zeta b_2a_3} + b_3a_2 \\ b_1a_2 + b_2a_1 + b_3a_3 \\ b_1a_3 + b_2a_2 + b_3a_1 \end{bmatrix} \tag{3.39}$$

We can infer constituent-order by comparing $\mathbf{m}_f$ and $\mathbf{m}_b$ as follows. First consider the dot product $\mathbf{m}_f \cdot \mathbf{m}_f$,

$$\begin{aligned} \mathbf{m}_f \cdot \mathbf{m}_f = (a_1b_1 + \zeta a_2b_3 + a_3b_2)^2 + \\ (a_1b_2 + a_2b_1 + a_3b_3)^2 + \\ (a_1b_3 + a_2b_2 + a_3b_1)^2 \end{aligned} \tag{3.40}$$

$$\begin{aligned} = (a_1^2b_1^2 + \zeta^2a_2^2b_3^2 + a_3^2b_2^2) + noise + \\ (a_1^2b_2^2 + a_2^2b_1^2 + a_3^2b_3^2) + noise + \\ (a_1^2b_3^2 + a_2^2b_2^2 + a_3^2b_1^2) + noise \end{aligned} \tag{3.41}$$

The expectation of this dot product can can be expressed as the sum of specific products between random variables. The noise terms can be dropped for expectation calculations because they consist of odd powers of random variables, and the expectation for odd powers of standard normal distributed variables is 0 (Anderson, 1970). Thus, the expectation of this dot product can be expressed as follows,

$$= \mathrm{E}\left[\mathbf{m}_f \cdot \mathbf{m}_f\right] = (n^2-1)\mathrm{E}\left[X^2Y^2\right] + \zeta^2\mathrm{E}\left[X^2Y^2\right] \tag{3.42}$$

Where X and Y denote random variables drawn from $N(0,\sigma^2)$. Following Weber (1988), expectations of squared random variables can be substituted,

$$= \mathrm{E}\left[\mathbf{m}_f \cdot \mathbf{m}_f\right] = (n^2-1)\sigma^4 + \zeta^2\sigma^4 \tag{3.43}$$

Assuming that each element from **a** and **b** is drawn from $N(0,\sigma^2)$, and $\sigma^2 = \frac{1}{n}$, which produces approximately normalized vectors, this equation can be simplified even further,

$$= \mathrm{E}\left[\mathbf{m}_f \cdot \mathbf{m}_f\right] = \frac{(n^2-1)+\zeta^2}{n^2} \tag{3.44}$$

Equation 3.44 reveals that $\mathrm{E}\left[\mathbf{m}_f \cdot \mathbf{m}_f\right]$ has a quadratic relationship to to ζ. For comparison, let us also derive $\mathrm{E}\left[\mathbf{m}_f \cdot \mathbf{m}_b\right]$, which is expanded as follows,

$$\begin{aligned} \mathbf{m}_f \cdot \mathbf{m}_b = {} & (a_1b_1+\zeta a_2b_3+a_3b_2)(b_1a_1+\zeta b_2a_3+b_3a_2)+ \\ & (a_1b_2+a_2b_1+a_3b_3)(b_1a_2+b_2a_1+b_3a_3)+ \\ & (a_1b_3+a_2b_2+a_3b_1)(b_1a_3+b_2a_2+b_3a_1) \end{aligned} \tag{3.45}$$

$$\begin{aligned} = {} & (a_1^2b_1^2+\zeta a_2^2b_3^2+\zeta a_3^2b_2^2)+noise+ \\ & (a_1^2b_2^2+a_2^2b_1^2+a_3^2b_3^2)+noise+ \\ & (a_1^2b_3^2+a_2^2b_2^2+a_3^2b_1^2)+noise \end{aligned} \tag{3.46}$$

Again dropping the noise terms because odd powers of random variables have expectations of zero, we can derive the expectation of this dot product as follows,

$$\mathrm{E}\left[\mathbf{m}_f \cdot \mathbf{m}_b\right] = (n^2-2)\mathrm{E}\left[X^2Y^2\right] + 2\zeta\mathrm{E}\left[X^2Y^2\right] \tag{3.47}$$

$$= (n^2-2)\sigma^4 + 2\zeta\sigma^4 \tag{3.48}$$

$$= \frac{(n^2-2)+2\zeta}{n^2} \tag{3.49}$$

These equations reveal that $\mathrm{E}\left[\mathbf{m}_f \cdot \mathbf{m}_b\right]$ has a linear relationship to ζ, while $\mathrm{E}\left[\mathbf{m}_f \cdot \mathbf{m}_f\right]$ has a quadratic relationship, meaning that the difference between these dot product would increase with ζ. Thus, in this extremely simple implementation of a neural network with unequal synaptic strengths (with only one "strong" synapse), we can start to see how this type of mechanism could introduce differences between the forward and backward versions of an item-item associations thatcould be leveraged to infer order, and without any explicit mechanism to encode order. In sum, the simple assumption that convolution is not be strictly commutative when implemented in the brain could provide a simple way to support order memory.

Applications to serial recall. Associations that are symmetric yet support some ability to discriminate the item position could also be useful for understanding how people remember ordered lists of items with serial recall. Associative chaining (e.g. Ebbinghaus, 1885/1913; Lewandowsky & Murdock, 1989) is a major class of model of serial recall, and assumes that a participant learns a list of words in order by forming direct item-item associations between neighbouring items. At test, the list is remembered by sequentially chaining through the items, using one item as the cue for its next. Although there is evidence for chaining-like effects (Caplan, 2015; Lindsey & Logan, 2019; Solway, Murdock, & Kahana, 2012), certain benchmark findings have argued against a pure chaining account. This has led to the proliferation of positional-coding models that strictly avoid inter-item associations, and associate each item with its own positional code (Conrad, 1960; Brown et al., 2007; Burgess & Hitch, 1999; Farrell, 2012; Henson, 1998), although even these models have trouble with different empirical findings (Solway et al., 2012). In sum, chaining cannot be completely ruled out (Caplan, 2015), but it may need some modifications to account for the latest benchmark findings, which is where our examined extensions to convolution may be useful.

Implementations of chaining such as Lewandowsky and Murdock (1989); Solway et al. (2012), and Caplan, Ardebili, and Liu (2022) have used symmetric operations like convolution to encode item-item associations. A symmetric chaining model matches well with certain behavioural benchmarks. For example, if participants accidentally skip an item during serial recall, participants frequently go backward in the list and recall the missed word (fill-in errors, Henson, 1998) rather than proceeding on with the list (in-fill errors).[8] Symmetric associations could be useful for a chaining model to account for these findings. If an item is skipped, (A,. . . C), having equal forward and backward association strengths would mean that using C as a cue would retrieve both B and D equally, and the model could produce fill-in errors with high frequency. Other findings that suggest item-item associations are somewhat directional. For example, cued recall of serial lists tends to exhibit forward asymmetries (Kahana & Caplan, 2002). Applying any our present

[8] Some studies have found that in-fill errors are more frequent (Solway et al., 2012; Caplan, 2015)

order-encoding mechanisms to a symmetric chaining model (e.g., left right patterns of partial permutation for each item-item association), would allow the model to have some ability to make use of the ability to discriminate the constituent-order of item-item association, while preserving the ability to progress backward and forward equally through the list. This type of model may be well-equipped to fit both benchmark findings that support symmetric associations, and findings that indicate some directionality to memory for serial lists.

3.5 Conclusion

Multiple modifications to convolution preserved important properties of this model while adding some ability to judge constituent-order. However, only position-specific permutations could successfully disambiguate double function pairs. This demonstrates that there are a number of possible mechanisms by which symmetry can co-exist with some ability to judge constituent-order, but the partial permutation model accounted for the broadest set of empirical benchmarks.

Chapter 4

Conclusions and synthesis, accounting for order and mental imagery within mathematical models

Our first research question was whether conscious experience of mental imagery and/or mental imagery skill was necessary to benefit from imagery instructions. We found that:

- Individual differences in both visual imagery vividness (Vividness of Visual Imagery Questionnaire ratings) and skill (Paper Folding Task performance) had null relationships to the effectiveness of imagery instructions
- Self-identified aphantasics who report little to no ability to form mental images exhibited no hint of reduced benefit from imagery instructions.

Our next research question was whether associations studied with visual imagery would be stored with more information for constituent-order. We found that:

- Interactive imagery instructions did not improve order recognition performance, nor change its moderate relationship with cued recall performance (chapter 2, experiment 1), even when participants were provided ways to incorporate order into the image (chapter 2, experiment 2).

Our third research question was whether existing models could be modified to store moderate levels of order. This was especially important to address given the robustness of moderate order memory to imagery instructions. We found the following:

- Convolution-based models, which otherwise have no ability to discriminate AB from BA, can be modified in four separate ways to discriminate order above-chance and capture the moderate relationship between order recognition and cued recall performance, without compromising the inherent symmetry of convolution.
- The additional constraint of double-function lists could only be satisfied by partial permutation variant of convolution (model Π).

In this final chapter, we elaborate on our discussion of these findings from earlier, which special focus on connections to mathematical models.

4.1 An alternative explanation of interactive imagery effects

Through the lens of dual-coding theory (Paivio, 1969, 1971, 1986) imagery instructions are effective because they elicit visual imagery, allowing participants to store both an imaginal and verbal format of the association. This explanation aligned with Kosslyn and colleagues' position in the imagery debate, which proposed that mental imagery corresponded to a distinct *depictive* format by which information is represented in the mind (Pearson & Kosslyn, 2015). If associations studied with interactive imagery are stored in both an imaginal and verbal format, this implies that individuals who can form more vivid, or accurate visual images should be able to store the visual format of the association with more detail, which could aid memory performance. Furthermore, individuals with little to no ability to form mental images (aphantasia) should be unable to store imaginal representations, and receive less benefits from imagery instructions.

However, individual differences in both imagery vividness and skill could not explain the effectiveness of interactive imagery. Furthermore, in self-identified aphantasics, we saw no reduced benefit from imagery instructions. These results suggest that the experience of mental imagery is not required for interactive imagery effects. This result was more consistent with the argument that visual imagery itself is epiphenomenal (Pylyshyn, 2002), and does not necessarily correspond to underlying cognitive mechanisms that lead memory to improve by a large margin. This motivated us to think about alternative ways by which interactive imagery instructions improve memory.

Although interactivity did not emerge from subjective reports in experiment 2 and 3 (see Chapter 2, Supplementary Materials), the ineffectiveness of top-bottom imagery compared to other instructions led us to suspect that imagining one word on top of the other, without any explicit instruction to imagine them interacting, might have experimentally reduced the rate of interactivity. Thus, we re-considered the potential benefits of imagining interactions between words. Our suggestion was that forming images, or non-visual analogues, which conceptualized explicit interactions between words in a pair could highlight item features that are pair-unique, which would have downstream benefits for association memory. To remind the reader of a helpful illustration, consider the different ways you would think about the word APPLE, if you saw it in OVEN APPLE versus TEACHER APPLE. Although there may be many ways one could implement this idea mathematically, one possible idea is concatenate additional features and increase total dimensionality of item vectors, representing the additional pair-unique details that the participant is led to think about after they receive interactive imagery instructions.

Future directions One way to test the validity of our present explanation of interactive imagery effects, is to modify the instructions to be even more effective based on our insights. Future studies could test instructions that remove references to mental imagery, which by our current findings do not seem essential, and add an emphasis on processing each word in a pair-unique way. For example, changing instructions in figure 3.2 from "Please try this technique for the next word pairs. Form a mental image with both of the words interacting together when you are presented with a word pair. For example, for the word pair CAT–DOG, you could imagine the cat chasing the dog." to "Please try this technique for the next word pairs. When you think about both of the words, think about them interacting together in a way that makes them unique to the pairing itself. For example, for the word pair OVEN APPLE, you could think about baked apples in a hot oven, while for the pair TEACHER APPLE, you could think about a happy teacher holding a polished red apple.”. If this strategy proves more effective, we might have more evidence for the importance of processing words in a way that is conducive to association memory performance.

4.2 Moderate memory for order is unaffected by imagery instructions

The initial idea that imagery instructions led participants to encode distinct visual or imaginal representations led us to hypothesize that imagery instructions might be an effective way to improve memory for order. We hypothesized that stored visual representations could provide participants with a better way to store and recover the constituent-order of associations, leading to increased order recognition performance, even to the maximal levels predicted by matrix models.

In experiment 1 in chapter 2, we found that interactive imagery instructions had no effect on order recognition performance, or its relationship to cued recall performance. In experiment 2 we modified standard imagery instructions to indicate how the participant was to encode order within the visual image, to address the possibility that participants did not improve because they could not think of an effective way to incorporate order into the image. These imagery instruction variants like actor-object imagery (e.g., DOG ate the PIE), seemed like a simple way to form an image incorporating both the association and its order. Yet, participants administered these variants exhibited no substantial benefit to order recognition performance. These results both replicate and extend Kato and Caplan (2017), demonstrating the surprising generality of mid-range order recognition performance, even when participants are given explicit study strategies that are quite effective to improve memory for the association itself.

As we discuss this later in this chapter, these results present an additional mystery about the nature of interactive imagery effects, and also about how participants represent the constituent-order of associations. Whatever effect that imagery instructions have to improve association memory is ineffective to improve order recognition, which must be accounted for in efforts to model or explain these effects.

Future directions Our results indicate that the moderate relationship between order recognition and cued recall performance is quite general, which has direct implications for the existing mathematical models. However, there may be certain factors that could influence the ability to store the constituent-order of associations, and further inform our understanding of order memory, as we discuss below.

First, in the experiments presented here, and in Kato and Caplan (2017), words in each pair were presented simultaneously. In this case, order judgements are specifically asking participants to remember the left-right spatial relationship between words within a pair. Another common presentation method is to present word pairs sequentially, where the first associate is presented on its own, followed shortly after by the second associate. In this case, judgments of constituent-order would ask participants to remember the temporal, first versus last, relationship between words. Some effects generalize between these two presentation methods. For example, J. Yang et al. (2013) and Madan, Glaholt, and Caplan (2010) found associative symmetry in both simultaneous and sequential presentation conditions. However, if we found that order recognition differed between these two conditions, it may suggest that models need to adopt different parameters in both conditions, or completely different models altogether for temporal and spatial order. Another parameter to consider is our presentation rate, which was 2.85 seconds/pair in all experiments. Had participants been given more time to incorporate order into formed images (or non-visual analogues) at study, this may have allowed them to improve order recognition performance post-imagery instruction. Future studies could test if order recognition performance and its relationship to association memory would be increased at longer presentation rates.

We should also consider that in many real-world scenarios we can make judgements about order based on our prior knowledge about stimuli. For example, with names such as BRIAN O'BRIEN, if you are familiar with both words, it is straightforward to judge which word is more likely first name, and which word is more likely last name. In this case, one can infer order based on prior knowledge about the individual words, rather than memory of a specific instance in which these words were seen together. Another example is compound words, which have modifier-head relationships (Dressler, 2006; Caplan, Boulton, & Gagné, 2014) where it is immediately clear what the correct order of the pair is based on prior knowledge (FISH HOOK versus HOOK FISH). These examples suggest that explicitly storing and retrieving the constituent-order of novel associations is often unnecessary, and may explain why we find that participants are considerably worse at judging order compared to remembering novel associations.

4.3 Producing ordered, symmetric associations with convolution models

One of the main theoretical developments in this thesis was the successful extension of symmetric convolution models to produce above-chance order recognition, and in a way that was moderately dependent on association memory. This was accomplished with minimal modifications that did not substantially increase model complexity, preserving useful characteristics of convolution such as its inherent symmetry. In chapter 3, we contrasted our attempts with previous attempts to modify matrix-based models to encode associations with *less* order (Kato & Caplan, 2017). These attempts proved more challenging, especially with the additional constraint of associative symmetry, although we discussed how our examined mechanisms may be applied to a symmetric matrix model to overcome previous issues. In any case, this analysis returned multiple ways that symmetric associations could also support moderate, behavioural levels of order memory, reconciling seemingly contradicting characteristics of verbal association memory.

We pushed our investigation even further by evaluating each of our model modifications against double function lists. We found that this additional constraint could only be satisfied by partial permutation (model Π). Item-position associations (model A), item and position vectors summation (model Σ) and position features (model ϕ) were all ruled out because position information was repeated for every item in the list, and did not provide any pair-specific information that could be used to solve AB versus BC interference. The success of permutation allowed us to make more specific conclusions about *how* order may be represented in memory, suggesting that order was encoded by modifying the item representation directly, rather than through associations to some separate position vector.

Future directions As we mentioned in chapter 3, an interesting limitation of model $\mathit{\Pi}$ and ϕ is that both models can only judge order above-chance if items are in their original pairing. Although our previous discussion of this model property was in regards to the correct-order advantage for associative recognition (J. Yang et al., 2013), predictions can also be generated about when the primary task is order recognition. For example, assume that model $\mathit{\Pi}$ encodes the following pairs, $\mathbf{m} = p_l(\mathbf{a}) * p_r(\mathbf{b}) + p_l(\mathbf{c}) * p_r(\mathbf{d})$. As we know, the model $\mathit{\Pi}$ can discriminate AB from BA, because $\mathrm{E}\left[p_l(\mathbf{a}) * p_r(\mathbf{b}) \cdot \mathbf{m}\right] > \mathrm{E}\left[p_r(\mathbf{a}) * p_l(\mathbf{b}) \cdot \mathbf{m}\right]$. However, the model cannot discriminate AD from DA, because $\mathrm{E}\left[p_l(\mathbf{a}) * p_r(\mathbf{d}) \cdot \mathbf{m}\right]$ and $\mathrm{E}\left[p_r(\mathbf{a}) * p_l(\mathbf{d}) \cdot \mathbf{m}\right]$ are both equal to 0. A similar argument applies to model ϕ, where $\mathrm{E}\left[(\mathbf{a} \oplus \mathbf{l}) * \mathbf{d} \oplus \mathbf{r} \cdot \mathbf{m}\right]$ and $\mathrm{E}\left[(\mathbf{a} \oplus \mathbf{r}) * \mathbf{d} \oplus \mathbf{l} \cdot \mathbf{m}\right]$ are also both equal to 0. In contrast, model A and $\mathit{\Sigma}$ can make independent judgements about the positions of items, by, for example, comparing $\mathrm{E}\left[(\mathbf{a} * \mathbf{l}) \cdot \mathbf{m}\right]$ and $\mathrm{E}\left[(\mathbf{a} * \mathbf{r}) \cdot \mathbf{m}\right]$.

These predictions could be tested empirically by asking participants to perform order recognition judgments for four types of probes AB, BA, AD and DA. Model ϕ and $\mathit{\Pi}$ could, as we know, produce above-chance order recognition d' for AB versus BA, but produce d' for AD versus DA should be 0. Models A and $\mathit{\Sigma}$ would predict comparable order recognition d' for AB versus BA, and AD versus DA. J. Yang et al. (2013) conducted an experiment that would test these predictions. The main focus of their study was associative recognition; however, J. Yang et al. (2013) also included a "directional judgement" task in Experiment 6 for pairs AB, BA, AD, and DA. Directional judgements here were essentially the same as our order recognition task, except that participants were instructed to judge if words were in their correct position regardless of whether in their correct pairing or not. For intact pairs, they found that *accuracy* for correct order pairs (AB) was significantly higher compared to accuracy for pairs in the incorrect order (BA), but there was no significant difference for recombined pairs between correct order (AD) and incorrect order (DA) pairs. These results seem consistent with model ϕ and $\mathit{\Pi}$; however, a direct evaluation of our models against this data would require more steps.

4.4 Common themes: accounting for imagery effects and order recognition data within a unified modelling framework

Throughout this thesis we have addressed two separate research questions somewhat independently. However, given our overall goal of understanding the computational characteristics of association memory, an interesting future direction may be to integrate imagery effects and our models of order memory. In chapter 2, we found that order recognition performance, and its correlation to cued recall performance, was unaffected by imagery instructions. This result indicates that whatever mechanism interactive imagery instructions are engaging to improve association memory, whether it be increased item features or something else entirely, is not effective to improve order memory. Besides emphasizing the need for models with mid-range order, this reveals an additional puzzle that future modelling work must address. How can imagery strategies provide detail that boosts association memory without improving order recognition?

We can derive some hints about whether each of our present extensions of convolution can accomplish this from previously reported simulations in chapter 3. If we focus on each model's "order parameter", parametric plots in figure 3.3 show that our models cannot independently modulate cued recall/associative recognition and order recognition performance. However, when models were able to vary parameters n and σ_α in figure 3.8 we see a different story. Models ϕ and Π could clearly produce a variety of cued recall accuracy values at the same order recognition d'. Models A and Σ produced a considerably narrower range of cued recall accuracy values at each value of order recognition d'; however, as we noted earlier, models A and Σ may be better at producing the same range in cued recall accuracy as the other models if they were fit to this data.

In any case, these simulations may indicate we already have models that can different combinations of free parameters to mimic imagery effects on association memory, without changing order recognition d'. Future work could test whether each of our models can account for these effects while closely examining how each model uses its parameters to do this.

4.5 Overall conclusions

In sum, we found no evidence for the idea that visual imagery is necessary for the benefits due to imagery instructions, consistent with existing arguments that the experience of imagery is epiphenomenal (Pylyshyn, 2002). This leaves open the possibility for alternative explanations of imagery effects.

In pursuit of our other major goal, we replicated and extended the finding that associations are remembered with moderate order, to conditions where participants are given imagery instructions. Additionally, we modified convolution-based models in multiple ways to resolve these challenges, although only the partial permutations (model Π) could solve the additional benchmark of double function lists.

Then, synthesizing the work here, we considered the additional finding that imagery instructions only benefit association memory and not order recognition performance. This presents an additional puzzle for future modelling efforts. Any account of imagery effects, or order memory, must also explain why the engaged mechanism after imagery instructions does not improve memory for order.

Bibliography

Allen, R. J., Waterman, A. H., Yang, T., & Jaroslawska, A. J. (2022). Working memory in action: Remembering and following instructions. In R. H. Logie, Z. Wen, S. E. Gathercole, N. Cowan, & R. W. Engle (Eds.), *Memory in science for society: There is nothing as practical as a good theory.* Oxford University Press.

Anderson, J. A. (1970). Two models for memory organization using interacting traces. *Mathematical Biosciences*, *8*, 137-160.

Asch, S. E., & Ebenholtz, S. M. (1962). The principle of associative symmetry. *Proceedings of the American Philosophical Society*, *106*(2), 135-163.

Bainbridge, W. A., Pounder, Z., Eardley, A. F., & Baker, C. I. (2021). Quantifying aphantasia through drawing: Those without visual imagery show deficits in object but not spatial memory. *Cortex*, *135*, 159-172.

Benjamin, A. S. (2010). Representational explanations of "process" dissociations in recognition: The dryad theory of aging and memory judgments. *Psychological Review*, *117*(4), 1055-1079.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice.* Cambridge, MA: MIT Press.

Borsellino, A., & Poggio, T. (1972). Holographic aspects of temporal memory and optomotor responses. *Kybernetik*, *10*(1), 58-60.

Bower, G. H. (1970a). Imagery as a relational organizer in associative learning. *Journal of Verbal Learning and Verbal Behavior*, *9*, 529-533.

Bower, G. H. (1970b). Organizational factors in memory. *Cognitive Psychology*, *1*, 18-46.

Bower, G. H., & Winzenz, D. (1970). Comparison of associative learning strategies. *Psychonomic Science*, *20*, 119-120.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433-436.

Bridges, D., Pitiot, A., MacAskill, M., & Pierce, J. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*.

Brooks, L. R. (1967). The supression of visualization by reading. *Quarterly Journal of Experimental Psychology*, *19*, 139-159.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539-576.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*(3), 551-581.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261-304.

Caplan, J. B. (2015). Order-memory and association-memory. *Canadian Journal of Experiemntal Psychology*, *69*(3), 221-232.

Caplan, J. B., Ardebili, A. S., & Liu, Y. S. (2022). Chaining models of serial recall can produce positional errors. *Journal of Mathematical Psychology*, *109*, 102677.

Caplan, J. B., Boulton, K. L., & Gagné, C. L. (2014). Associative asymmetry of compound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 1163-1171.

Caplan, J. B., Chakravarty, S., & Dittmann, N. L. (2021). Associative recognition without hippocampal associations. *Psychological Review*.

Caplan, J. B., & Madan, C. R. (2016). Word-imageability enhances association-memory by increasing hippocampal engagement. *Journal of Cognitive Neuroscience*, *28*(10), 1522-1538.

Caplan, J. B., Rehani, M., & Andrews, J. C. (2014). Associations compete directly in memory. *Quarterly Journal of Experimental Psychology*, *67*(5), 955-978.

Conrad, R. (1960). Serial order intrusions in immediate memory. *British Journal of Psychology*, *51*(1), 45-48.

Cox, G. E., & Criss, A. H. (2017). Parallel interactive retrieval of item and associative information from event memory. *Cognitive Psychology*, *97*(5), 31–61.

Cox, G. E., & Criss, A. H. (2020). Similarity leads to correlated processing: A dynamic model of encoding and recognition of episodic associations. *Psychological Review*, *127*(5), 792–828.

Cox, G. E., Hemmer, P., Aue, W. R., & Criss, A. H. (2018). Information and processes underlying semantic and episodic memory across tasks, items, and individuals. *Journal of Experimental Psychology: General*, *147*(4), 545-590.

Criss, A. H., & Shiffrin, R. M. (2004a). Interactions between study task, study time, and the low-frequency hit rate advantage in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 778-786.

Criss, A. H., & Shiffrin, R. M. (2004b). Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory & Cognition*, *32*, 1284–1297.

Criss, A. H., & Shiffrin, R. M. (2005). List discrimination in associative recognition and implications for representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1199-1212.

de Beni, R., & Cornoldi, C. (1985). The effects of imaginal mnemonics on congenitally total blind and on normal subjects. In D. F. Marks & D. Russell (Eds.), *Imagery 1* (p. 56-59). Dunedin, N.Z.: Human Performance Associates.

de Vito, S., & Bartolomeo, P. (2016). Refusing to imagine? On the possibility of psychogenic aphantasia. A commentary on Zeman et al. (2015). *Cortex*, 334-335.

Dempster, F. N., & Rohwer, W. D. (1974). Component analysis of the elaborative encoding effect in paired-associate learning. *Journal of Experimental Psychology*, *103*(3), 400-408.

Dressler, W. U. (2006). Compound types. In G. Libben & G. Jarema (Eds.), *The representation and processing of compound words* (p. 23-44). Oxford University Press.

Duncan, K., Tompary, A., & Davachi, L. (2014). Associative encoding and retrieval are predicted by functional connectivity in distinct hippocampal area CA1 pathways. *Journal of Neuroscience*, *34*(34), 11188-11198.

Dunlosky, J., Hertzog, C., & Powell-Moman, A. (2005). The contribution of mediator-based deficiencies to age differences in associative learning. *Developmental Psychology*, *41*(2), 389-400.

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology*. New York: Teachers College, Columbia University.

Engelkamp, J. (1991). Imagery and enactment in paired-associate learning. In R. H. Logie & D. M. (Eds.), *Mental images in human cognition* (p. 119-128). Amsterdam: North Holland Press.

Engelkamp, J. (1995). Visual imagery and enactment of actions in memory. *British Journal of Psychology*, *86*, 227-240.

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*(2), 223-271.

Faw, B. (2009). Conflicting intuitions may be based on differing abilities. *Journal of Conciousness Studies*, *16*(4), 45-68.

Fiebach, C. J., & Friederici, A. D. (2004). Processing concrete words: fMRI evidence against a specific right-hemisphere involvement. *Neuropsychologia*, *42*(1), 62-70.

Foer, J. (2011). *Moonwalking with Einstein: The Art and Science of Remembering Everything*. New York, NY: Penguin Press.

Fraundorf, S. H., Diaz, M., Finley, J., Lewis, M. L., Tooley, K. M., Isaacs, A. M., ... Brehm, L. (2014). *CogToolbox for MATLAB [computer software].* Retrieved from `http://www.scottfraundorf.com/cogtoolbox.html`

French, J. W., Ekstrom, R. B., & Price, L. A. (1963). *Kit of reference tests for cognitive factors*. Princeton, NJ: Educational Testing Service.

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods, Instruments, & Computers*, *14*, 375–399.

Gabor, D. (1969). Associative holographic memories. *IBM Journal of Research and Development*, *13*(2), 156-159.

Geller, A. S., Schleifer, I. K., Sederberg, P. B., Jacobs, J., & Kahana, M. J. (2007). PyEPL: a cross-platform experiment-programming library. *Behavior Research Methods*, *39*(4), 950-958.

Gesualdo, F. (1592). *Plutosofia*. Padua.

Giovanello, K. S., Schnyer, D., & Verfaellie, M. (2009). Distinct hippocampal regions make unique contributions to relational memory. *Hippocampus*, *19*, 111-117.

Greene, R. L., & Tussing, A. A. (2001). Similarity and associative recognition. *Journal of Memory and Language*, *45*, 573-584.

Haskins, A. L., Yonelinas, A. P., Quamme, J. R., & Ranganath, C. (2008). Perirhinal cortex supports encoding and familiarity-based recognition of novel associations. *Neuron*, *59*, 554-560.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46-51.

Henson, R. N. A. (1998). Short-term memory for serial order: the Start-End Model. *Cognitive Psychology*, *36*(2), 73-137.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96-101.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528-551.

Hockley, W. E. (1994). Reflections of the mirror effect for item and associative recognition. *Memory & Cognition*, *22*(6), 713-722.

Hockley, W. E., & Cristi, C. (1996). Tests of encoding tradeoffs between item and associative information. *Memory & Cognition*, *24*, 202-216.

Horowitz, L. M., Brown, Z. M., & Weissbluth, S. (1964). Availability and the direction of associations. *Journal of Experimental Psychology*, *68*(6), 541-549.

Howard, M. W., Jing, B., Rao, V. A., Provyn, J. P., & Datey, A. V. (2009). Bridging the gap: transitive associations between items presented in similar temporal contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 391-407.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 923-941.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*(2), 208-233.

JASP Team. (2021). *JASP (Version 0.15)[computer software].* Retrieved from `https://jasp-stats.org/[jasp-stats.org]`

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1-37.

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition*, *30*(6), 823-840.

Kahana, M. J. (2012). *Foundations of Human Memory*. USA: Oxford University Press.

Kahana, M. J., & Caplan, J. B. (2002). Associative asymmetry in probed recall of serial lists. *Memory & Cognition*, *30*(6), 841-849.

Kato, K., & Caplan, J. B. (2017). Order of items within associations. *Journal of Memory and Language*, *97*, 81-102.

Kelly, M. A., Blostein, D., & Mewhort, D. J. K. (2013). Encoding structure in holographic reduced representations. *Canadian Journal of Experimental Psychology*, *67*(2), 79-93.

Kelly, M. A., Mewhort, D. J. K., & West, R. L. (2017). The memory tesseract: mathematical equivalence between composite and separate storage memory models. *Journal of Mathematical Psychology*, *77*, 142-155.

Keogh, R., & Pearson, J. (2018). The blind mind: No sensory visual imagery in aphantasia. *Cortex*, *7*, 53-80.

Klaver, P., Fell, J., Dietl, T., Schür, S., Schaller, C., Elger, C. E., & Fernández, G. (2005). Word imageability affects the hippocampus in recognition memory. *Hippocampus*, *15*, 704-712.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in psychtoolbox-3? *Perception*, *36*(14), 1-16.

Kliegl, R., Smith, J., & Baltes, P. B. (1990). On the locus and process of magnification of age differences during mnemonic training. *Developmental Psychology*, *26*, 894-904.

Kluger, F. E., Oladimeji, D. M., Tan, Y., Brown, N. R., & Caplan, J. B. (2022). Mnemonic scaffolds vary in effectiveness for serial recall. *Memory*.

Konrad, B. N. (2013). *Superhirn - Gedächtnistraining mit einem Weltmeister*. Vienna: Goldegg Verlag.

Kounios, J., Bachman, P., Casasanto, D., Grossman, M., & Smith, W., Roderick W. Yang. (2003). Novel concepts mediate word retrieval from human episodic associative memory: evidence from event-related potentials. *Neuroscience Letters*, *345*, 157-160.

Kounios, J., Smith, R. W., Yang, W., Bachman, P., & D'Esposito, M. (2001). Cognitive association formation in human memory revealed by spatiotemporal brain imaging. *Neuron*, *29*, 297-306.

Kucera, H., & Francis, W. (1967). Computational analysis of present-day American English. Providence, R.I.: Brown University Press.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (p. 112-146). New York, NY: John Wiley and Sons.

Lewandowsky, S., & Murdock, B. B. (1989). Memory for serial order. *Psychological Review*, *96*(1), 25-57.

Lindsey, D. R., & Logan, G. D. (2019). Item-to-item associations in typing: Evidence from spin list sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(3), 397-416.

Longuet-Higgins, H. C. (1968). Holographic model of temporal recall. *Nature*, *217*, 104.

Madan, C. R., Glaholt, M. G., & Caplan, J. B. (2010). The influence of item properties on association-memory. *Journal of Memory and Language*, *63*(1), 46-63.

Maguire, E. A., Valentine, E. R., Wilding, J. M., & Kapur, N. (2003). Routes to remembering: the brains behind superior memory. *Nature Neuroscience*, *6*(1), 90-95.

Marks, D. F. (1972). Individual Differences in the Vividness of Visual Imagery and Their Effect on Function. In P. W. Sheehan (Ed.), *The Function and Nature of Imagery* (p. 83-106). New York: Academic Press.

Marks, D. F. (1973). Visual imagery differences in the recall of pictures. *British Journal of Psychology*, *64*(1), 17-24.

McKelvie, S. J. (1995). The VVIQ as a psychometric test of individual differences in visual imagery vividness: A critical quantitative review and plea for direction. *Journal of Mental Imagery*, *19*(3-4), 1-106.

Metcalfe Eich, J. (1982). A composite holographic associative recall model. *Psychological Review*, *89*(6), 627-661.

Müller, N. C. J., Konrad, B. N., Kohn, N., Muñoz-López, M., Czisch, M., Fernández, G., & Dresler, M. (2018). Hippocampal–caudate nucleus interactions support exceptional memory performance. *Brain Structure and Function*, *223*(3), 1379–1389.

Murdock, B. B. (1962). Direction of recall in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *1*(2), 119-124.

Murdock, B. B. (1974). *Human memory: Theory and data.* Potomac, MD: Lawrence Erlbaum and Associates.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(6), 609-626.

Murdock, B. B. (1985). Convolution and matrix systems: a reply to Pike. *Psychological Review*, *92*(1), 130-132.

Murdock, B. B. (1995). Developing TODAM: three models for serial-order information. *Memory & Cognition*, *23*(5), 631-645.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251-269.

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, *76*(3), 241-263.

Paivio, A. (1971). *Imagery and Verbal Processes.* New York: Holt, Rinehart and Winston, Inc.

Paivio, A. (1986). *Mental Representations: A Dual Coding Approach.* New York: Oxford University Press.

Paivio, A., & Foth, D. (1970). Imaginal and verbal mediators and noun concreteness in paired-associate learning: The elusive interaction. *Journal of Verbal Learning and Verbal Behavior*, *9*, 384-390.

Paivio, A., Smythe, P. C., & Yuille, J. C. (1968). Imagery versus meaningfulness of nouns in paired associate learning. *Canadian Journal of Psychology*, *22*, 427-441.

Paivio, A., & Yuille, J. C. (1969). Changes in associative strategies and paired-associate learning over trials as a function of word imagery and type of learning set. *Journal of Experimental Psychology*, *79*(3), 458-463.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery and meaningfulness values for 925 words. *Journal of Experimental Psychology*, *76*(1, Part 2 Supplement), 1-25.

Paivio, A., Yuille, J. C., & Smythe, P. C. (1966). Stimulus and response abstractness, imagery, and meaningfulness, and reported mediators in paired-associate learning. *Canadian Journal of Psychology*, *20*(4), 362-377.

Pearson, J., & Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the National Academy of Sciences, USA*, *112*(33), 10089–10092.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195-203.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437-442.

Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review*, *91*(3), 281-294.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*(3), 623-641.

Plate, T. A. (2000). Analogy retrieval and processing with distributed vector representations. *Expert Systems: The International Journal of Knowledge Engineering and Neural Networks*, *17*(1), 29-40.

Pribram, K. H. (1969). The neurophysiology of remembering. *Scientific American*, *220*, 73-86.

Primoff, E. (1938). Backward and forward associations as an organizing act in serial and in paired-associate learning. *Journal of Psychology*, *5*, 375-395.

Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and Brain Sciences*, *25*, 157–238.

Recchia, G., Jones, M. N., Sahlgren, M., & Kanerva, P. (2010). Encoding sequential information in vector space models of semantics: comparing holographic reduced representation and random permutation. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd cognitive science society* (p. 865-870).

Rehani, M., & Caplan, J. B. (2011). Interference and the representation of order within associations. *Quarterly Journal of Experimental Psychology*, *64*(7), 1409-1429.

Richardson, J. T. E. (1985). Converging operations and reported mediators in the investigation of mental imagery. *British Journal of Psychology*, *75*, 205-214.

Richardson, J. T. E. (1998). The availability and effectiveness of reported mediators in associative learning: A historical review and an experimental investigation. *Psychonomic Bulletin & Review*, *5*(4), 597-614.

Rizzuto, D. S., & Kahana, M. J. (2000). Associative symmetry vs. independent associations. *NeuroComputing*, *32-33*, 973-978.

Rizzuto, D. S., & Kahana, M. J. (2001). An autoassociative neural network model of paired-associate learning. *Neural Computation*, *13*, 2075-2092.

Sanchez, C. (2019). The utility of visuospatial mnemonics is dependent on visuospatial aptitudes. *Applied Cognitive Psychology*, *33*, 519-529.

Shepard, R. N., & Metzler, J. (1973). Mental rotation of three-dimensional objects. *Science*, *171*, 701–703.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM— Retrieving Effectively From Memory. *Psychonomic Bulletin & Review*, *4*, 145-166.

Sivashankar, Y., & Fernandes, M. A. (2021). Enhancing memory using enactment: does meaning matter in action production? *Memory*, *30*, 147-160.

Slamecka, N. J. (1976). An analysis of double-function lists. *Memory & Cognition*, *4*(5), 581-585.

Solway, A., Murdock, B. B., & Kahana, M. J. (2012). Positional and temporal clustering in serial order memory. *Memory & Cognition*, *40*(2), 177-190.

Sommer, T., Schoell, E., & Büchel, C. (2008). Associative symmetry of the memory for object–location associations as revealed by the testing effect. *Acta Psychologica*, *128*, 238-248.

Thomas, J. J., Ayuno, K. C., Kluger, F. E., & Caplan, J. B. (in press). The relationship between interactive-imagery instructions and association-memory. *Memory & Cognition*.

van Heerden, P. J. (1963). A new optical method of storing and retrieving information. *Applied Optics*, *2*(4), 387-392.

Vincente, K. J., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review*, *105*(1), 33-57.

Weber, E. U. (1988). Expectation and variance of item resemblance distributions in a convolution-correlation model of distributed memory. *Journal of Mathematical Psychology*, *32*, 1-43.

Westbury, C. F., Cribben, I., & Cummine, J. (2016). Imaging imageability: Behavioral effects and neural correlates of its interaction with affect and context. *Frontiers in Human Neuroscience*, *10*, 346.

Westbury, C. F., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2013). Now you see it, now you don't: on emotion, context, and the algorithmic prediction of human imageability judgments. *Frontiers in Psychology*, *4*(991), 1-13.

Wiegand, I., Bader, R., & Mecklinger, A. (2010). Multiple ways to the prior occurrence of an event: an electrophysiological dissociation of experimental and conceptually driven familiarity in recognition memory. *Brain Research*, *1360*, 106-118.

Yang, J., Zhao, P., Zhu, Z., Mecklinger, A., Fang, Z., & Han, L. (2013). Memory asymmetry of forward and backward associations in recognition tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 253-269.

Yang, T., Allen, R. J., Waterman, A. H., Zhang, S., Su, X., & Chan, R. C. K. (2021). Comparing motor imagery and verbal rehearsal strategies in children's ability to follow spoken instructions. *Journal of Experimental Child Psychology*, *203*, 105033.

Yates, F. A. (1966). *The Art of Memory*. Chicago: University of Chicago Press.

Zeman, A., Della Sala, S., Torrens, L. A., Gountouna, V.-E., McGonigle, D. J., & Logie, R. H. (2010). Loss of imagery phenomenology with intact visuo-spatial task performance: A case of ‘blind imagination’. *Neuropsychologia*, *48*, 145-155.

Zeman, A., Dewar, M., & Della Sala, S. (2015). Lives without imagery - congenital aphantasia. *Cortex*, *73*, 378-380.

Zeman, A., Milton, F., Della Sala, S., Dewar, M., Frayling, T., Gaddum, J., . . . Winlove, C. (2020). Phantasia - the psychological significance of lifelong visual imagery vividness extremes. *Cortex*, *130*, 426-440.