Interference effects from machine learning matched confusable object pairs in memory assessment tasks

A thesis submitted for the degree of
Doctor of Philosophy

Devesh Batra
Hertford College
University of Oxford
Hilary 2022

Abstract

For early detection of neurodegeneration, objective cognitive tests provide a capable alternative to costly and invasive traditional biomarkers, as well as to short and infrequent in-clinic cognitive assessments. These tests, that are designed to target the cognitive processes first vulnerable to change in AD, can be remotely deployed in individuals' lives over long durations for frequent and repeated measurements using widely accessible digital devices such as smartphones. A class of such objective tests loads on the susceptibility of the short-term memory to interference from perceptually similar objects to obtain the first signal of AD-specific neurodegeneration. Evidence suggests that increasing the similarity of the confusable object pairs used in such tasks could have differential impact on the ability of individuals at different stages of cognitive health to inhibit interference from such pairs, thereby making object pair similarity a useful metric in sensitively recording individual cognitive differences. While this claim has been tested across various stimulus pair types, due to the lack of standardised datasets of confusable object pairs parametrically varying in similarity where the constituent objects in the pairs differ conceptually but still share some degree of similarity (e.g., *guitar* and *banjo*), such an assessment remains unexplored for this stimulus type. In this thesis, this gap is addressed by first using machine learning based models of vision and semantics, as well as their combinations, to produce the required datasets of object pairs parametrically varying in similarity, followed by exploring the differential effects of such confusable object pairs on interference in short-term memory among participants from varied demographics in an objective test of cognition.

First, an exploratory review of literature was conducted to study the methodologies that exist for producing datasets of confusable image pairs parametrically varying in similarity across stimulus types. The reviewed evidence suggested that while methods to produce such datasets exist for stimulus types such as face stimuli, abstract stimuli, and natural scene stimuli, such methods were lacking for generating confusable object pairs central to this thesis. In fact, a huge reliance on human input to generate such pairs was found. During this review, the relationship between stimulus pair similarity and size of interference was also noted. Studies showed a direct relationship between stimulus pair similarity and the magnitude of interference they produced. Last but not least, the interaction of interference was also examined by age. Studies unanimously agreed that older population is more susceptible to interference from confusable stimuli than younger population.

To address the gaps identified by the review, I used computer vision-based, linguistic association-based and taxonomic information-based computational models, as well as their combinations, to generate similar object pairs parametrically varying in similarity. To measure the ability of such computational models at estimating the human understanding of object pair similarity, the employed models as well as the datasets of object pairs they produced were rated against existing human-rated databases of object pair similarities. The object pairs produced by the best performing models from this analysis were then rated by 554 humans using a similarity rating experiment hosted on the Amazon mechanical Turk platform. Finally, these validated object pairs were deployed in Mezurio Gallery Blast, an objective cognitive test of short-term memory, to assess the utility of such pairs in the assessment of cognition.

The findings from the analysis of the performance of 1,455 cognitively normal participants who attempted the task as part of a remote study revealed that 1) the computationally matched object pairs employed in the task were capable of stimulating interference effects across participants in both the proactive and retroactive conditions of interference, 2) the object pair similarity has a significant and direct effect on susceptibility to retroactive interference but not to proactive interference, and 3) increasing object pair similarity in the retroactive interference condition has a larger effect on the older age group's susceptibility to interference compared to that of the younger age group. These results establish the utility of the computationally matched object pairs and the associated object pair similarity metric, in appropriately designed smartphone-based cognitive tasks. Furthermore, they provide preliminary evidence that such object pairing may facilitate detection of AD at an earlier stage of progression, and thus motivate future work.

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Oxford. This thesis contains approximately 50,000 words.

Acknowledgements

There are many people I am indebted to, without whom this thesis might not have been in the form it is.

First, I would like to thank my supervisor, Chris Hinds, who provided me with this opportunity to fulfil my dream of pursuing a research project in digital health. I would also like to thank him for his patience and guidance through the years that have (hopefully) made me a better researcher.

My thanks also go:

To all the volunteers who kindly gave up their time to take part in our research, especially the GameChanger study.

To John Gallacher, Richard Everson, and Christoffer Nellåker, for their valuable feedback in my vivas through my DPhil journey.

To Claire Lancaster, my favourite lab-mate, for the many useful discussions about cognitive tasks, data analyses and thesis writing, especially during the COVID period, that always helped progress my dissertation.

To my lab-mates, Simon Bond, for answering my umpteen questions about the Mezurio app, the GameChanger data and the idiosyncrasies of the British culture, and Alankar Atreya, for the many chats on machine learning as well as life as a DPhil student.

To the BDI Imaging Reading Group, for the very useful discussions on the latest in machine learning, that kept me up to date with the field.

To my always-at-the-BDI friends, Korsuk, Shing and Hamid, for the many insightful discussions through the years, especially during the pandemic, and my extended lab-mates at the BDI, Andreas, Anna, Dan, Mei and Soroosh, for the many coffee chats, evening pub trips, and endless banter.

To my friends in Oxford, Deepanshu, Vishal, Alex, Madhav, Dharam, Rebecca, Saumya, Miguel, Rayner, and Aayush, who never made me feel away from home.

To Linda, for sharing the thesis' ups and downs, for patiently listening to my various data analysis ideas, and for constantly showing me that there is more to life than beta coefficients. Also, to Sylvia and Helmar Gerlach, for reminding me to take breaks and unwind during my thesis writing phase.

Finally, and most importantly, to my family - my mother, Rma Batra, for always pushing me to pursue excellence, my father, Man Mohan Batra, for instilling in me the values of hard work and perseverance, and my little sister, Divya Batra, for being my constant partner in my endeavours. Thank you for your patience and encouragement throughout my thesis writing, for always cheering for me, and for the many virtual hugs you sent me through the COVID-19 pandemic from 4500 miles away.

Table of Contents

ABSTRACT **2**

ACKNOWLEDGEMENTS **4**

TABLE OF CONTENTS **5**

ABBREVIATIONS **9**

LIST OF FIGURES **10**

LIST OF TABLES **13**

CHAPTER 1: INTRODUCTION **14**

1.1 NEED FOR NEW TOOLS FOR EARLY DETECTION OF AD 14
1.2 EARLIEST CORTICAL SITE IN AD PATHOLOGY 15
1.3 COGNITIVE TASKS TARGETING PRC FUNCTION 17
1.4 CASE STUDY: MEZURIO GALLERY GAME 18
1.4.1 Task Design *19*
1.4.2 Confusability generation scheme *20*
1.4.3 Findings *21*
1.5 NEED FOR FINE-GRAINED ASSESSMENT OF CONFUSABILITY 22
1.6 CHARACTERISTICS OF SIMILAR PAIR DATASETS 23
1.7 FEATURE NORMS FOR MEASURING OBJECT PAIR SIMILARITY 26
1.8 MOTIVATION AND AIM OF PRESENT RESEARCH 28
1.9 THESIS STRUCTURE 30

CHAPTER 2: BACKGROUND: CONFUSABILITY GENERATION SCHEMES IN MEMORY ASSESSMENT TASKS **32**

2.1 FINE-GRAINED METRICS OF CONFUSABILITY IN MEMORY ASSESSMENT TASKS 33
2.1.1 Fine-grained assessment using natural scene stimuli *34*
2.1.2 Fine-grained assessment using abstract image stimuli *36*
2.1.3 Fine-grained assessment using everyday object pairs (related at basic category level) *40*
2.1.4 Key inferences *44*
2.2 CONFUSABILITY GENERATION WITH OBJECT PAIRS RELATED AT SUPERORDINATE CATEGORY LEVEL 44
2.2.1 Binary measure of confusability *46*
2.2.2 Granular assessment of similarity between distinct object pairs *48*
2.2.3 Observations *49*
2.3 AUTOMATED MODELS OF OBJECT REPRESENTATION 51

2.4 Discussion 53
2.5 Conclusion 54
CHAPTER 3: BACKGROUND: COMPUTATIONAL RESOURCES AND TECHNIQUES 55
3.1 Introduction 55
3.2 Computational models of object pair similarity 55
3.2.1 Distributional semantic models for linguistic similarity 56
3.2.2 Deep convolutional neural networks for visual similarity 61
3.2.3 Lexical Semantic Networks for taxonomic similarity 64
3.2.4 Similarity measures for vector space models 66
3.3 Production of confusable object pairs: The Hungarian algorithm 70
3.4 Evaluation techniques 72
3.4.1 Assessment on semantic category membership 72
3.4.2 Comparison with human judgements of concept pair similarity 75
3.5 Conclusion 77
CHAPTER 4: PRODUCTION OF CONFUSABLE OBJECT PAIRS USING COMPUTATIONAL MODELS 79
4.1 Introduction 79
4.2 Acquisition of the object image database 80
4.2.1 Past methods for acquiring object images 80
4.2.2 Criteria for image selection 82
4.2.3 Acquisition of images 83
4.2.4 The assembled image database 88
4.3 Production of confusable object pairs using computational models 89
4.3.1 Linguistic similarity based confusable pair production 89
4.3.2 Visual similarity based confusable pair production 90
4.3.3 Taxonomic similarity based confusable pair production 91
4.3.4 Bimodal visuo-linguistic similarity based confusable pair production 92
4.3.5 Bimodal visuo-taxonomic similarity based confusable pair production 97
4.3.6 Retrofitted visuo-linguistic similarity based confusable pair production. 100
4.4 Analysis 105
4.5 Results 106
4.5.1 Model performance on semantic category membership 107
4.5.2 Model performance with human judgements of object pair similarity 108
4.6 Discussion 108
4.6.1 Comparison among the unimodal models 109
4.6.2 Bimodal vs. unimodal models 112
4.6.3 Wordnet-retrofitted visuo-linguistic model 112
4.7 Conclusion 113

CHAPTER 5: IMSIM-1498: A DATASET OF HUMAN-RATED SIMILARITY VALUES FOR COMPUTATIONALLY MATCHED OBJECT PAIRS 115

5.1 INTRODUCTION 115
5.2 SIMILARITY RATING QUESTIONNAIRE DEVELOPMENT 117
5.2.1 Key traits of existing similarity rating questionnaires 117
5.2.2 New similarity rating task format 121
5.3 MATERIALS AND METHODS 122
5.3.1 Participants 122
5.3.2 Stimuli 124
5.3.3 Task design 124
5.3.3.1 Qualifying test 124
5.3.3.2 Questionnaire group design 127
5.3.3.3 Questionnaire group split 130
5.3.4 Post processing 131
5.4 STATISTICAL ANALYSIS 133
5.5 RESULTS 135
5.5.1 Inter-annotator agreement 135
5.5.2 Evaluation of computational models against ImSim-1498 136
5.5.3 Evaluation of the usability of computationally generated similarity values 138
5.6 DISCUSSION 139
5.7 CONCLUSION 143

CHAPTER 6: ASSESSING THE UTILITY OF COMPUTATIONALLY MATCHED OBJECT PAIRS IN AN OBJECTIVE COGNITIVE TASK 144

6.1 INTRODUCTION 144
6.2 MATERIALS AND METHODS 148
6.2.1 Participants 148
6.2.2 Procedure 150
6.2.3 Stimuli 151
6.2.4 Mezurio Gallery Blast 152
6.2.4.1 Task Description and outcome variables 153
6.2.5 Variables of Interest 158
6.2.5.1 Demographic-level 159
6.2.5.2 Stimulus-level 161
6.2.6 Participant exclusion 162
6.3 STATISTICAL ANALYSIS 165
6.4 RESULTS 168
6.4.1 Computationally produced confusable pairs inflict interference effects. 168
6.4.2 Susceptibility to both PI and RI increases with age 170

6.4.3 The effect of initial learning on errors due to both PI and RI decreases with age 170
6.4.4 Susceptibility to both PI and RI decreases with higher maximum education attained 171
6.4.5 Susceptibility to RI but not PI increases with increase in object pair similarity..... 171
6.4.6 Effect of object pair similarity increases with age in the RI but not the PI condition 172
6.4.7 Stimulus-level covariate of no interest 173
6.4.8 Other demographic-level covariates..... 174
6.5 DISCUSSION 175
6.5.1 Examination of the validity of the Gallery Blast task 175
6.5.1.1 Relationship between interference and age..... 176
6.5.1.2 Relationship between interference and other demographic-level variables 177
6.5.2 Potential of the computationally generated confusable object pairs in inducing interference in Gallery Blast 181
6.5.3 Differential effects of the object pair similarity metric on PI and RI 181
6.5.4 Differential effects of the object pair similarity metric on different age groups..... 183
6.5.5 Potential utility of the object pair similarity metric in detection of impairment 184
6.6 LIMITATIONS 187
6.7 CONCLUSION 187
CHAPTER 7: CONCLUSION..... 190
7.1 SUMMARY OF CONTRIBUTIONS 196
7.2 FUTURE DIRECTIONS 199
7.3 CONCLUDING REMARKS 199
APPENDIX A: WordNet Synsets and definitions of semantic categories . 201
APPENDIX B: List of all concepts and their WordNet Synsets..... 203
REFERENCES..... 246

Abbreviations

AD	Alzheimer's Dementia
API	Application Programming Interface
BNC	British National Corpus
CNN	Convolutional Neural Network
DMS	Delayed matching-to-sample
EDoN	Early Detection of Neurodegenerative Diseases
ERc	Entorhinal cortex
fMRI	Functional Magnetic Resonance Imaging
ILSVR	ImageNet Large Scale Visual Recognition
LSO	Lowest superordinate
MTL	Medial temporal lobe
OR	Odds Ratios
PI	Proactive Interference
PRc	Perirhinal cortex
RADAR-AD	Remote Assessment of Disease and Relapse – Alzheimer's Disease
RI	Retroactive Interference
SCD	Subjective Cognitive Decline
SD	Standard Deviation
VAS	Visual Analogue Scale
VSM	Vector Space Model
WUP	Wu and Palmer similarity metric

List of Figures

1.1 Lateral view of the human cerebral cortex demonstrating the ventral visual stream showing object processing pathway according to the representational-hierarchical theory.

1.2 Representation of learning, recognition and recall trials in Gallery Game.

1.3 Examples of target-distractor pairs in Gallery Game.

1.4 Schematic diagram depicting scope of the thesis.

3.1 Diagrammatic representation of linguistic vector space models in two dimensions.

3.2 Example of the working of skip-gram architecture used by Word2Vec.

3.3 Example of linguistic feature vectors stored row-wise in a matrix.

3.4 Visualisation of progressive encoding of image features in deep Convolutional Neural Networks.

3.5 Architecture of the Inception-v3 deep Convolutional Neural Network model.

3.6 Example WordNet sub-tree.

3.7 Example similarities between two concepts.

3.8 Depiction of the Hungarian algorithm.

4.1 Line drawn pictures from a past image stimuli dataset.

4.2 Pictures for the concept *'pineapple'* from ImageNet and Shutterstock.

4.3 Pipeline for image database acquisition.

4.4 Examples of unsuitable images for the present memory assessment task.

4.5 Google n-gram frequencies of concept occurrence as a proxy for familiarity.

4.6 Example images from the acquired dataset in this thesis.

4.7 Frequencies of concepts by semantic category in the image database from this thesis.

4.8 Pipeline for generating concept pairs using the linguistic model.

4.9 Pipeline for generating concept pairs using the visual model.

4.10 Pipeline for generating concept pairs using the taxonomic model.

4.11 Pipeline for generating concept pairs using the visuo-linguistic model.

4.12 Spearman's rank correlation of models produced using equation 4.1 with SimLex-999, SemSim and VisSim, for $\alpha \in [0,1]$.

4.13 Pipeline for generating concept pairs using the visuo-taxonomic model.

4.14 Spearman's rank correlation of models produced using equation 4.3 with SimLex-999, SemSim and VisSim for $\alpha \in [0,1]$.

4.15 Pipeline for generating concept pairs using the retrofitted visuo-linguistic model with 20 neighbours.

4.16 Spearman's rank correlation of models produced using equation 4.3 with SemSim and VisSim for $\alpha \in [0,1]$.

4.17 Percentage pairs with shared semantic categories produced by each model.

4.18 Performance of computational models on standard evaluation datasets.

4.19 Matched pairs produced by the linguistic, visual, and taxonomic models.

4.20 Qualitative example of improvement in deep-CNN model performance upon grounding in semantic information.

5.1 Examples used to train participants in the Amazon mechanical Turk study.

5.2 Mandatory qualifying test in the Amazon mechanical Turk study.

5.3 Example groups rated by the participants in the main study.

5.4 Visual Analogue Scale used in the Amazon mechanical Turk study.

5.5 Example images from the new dataset in increasing order of rated similarity.

5.6 Distribution of image pairs by ImSim-1498 similarity ratings.

5.7 Inter-annotator correlation on similarity rating datasets.

5.8 Change in the rankings of computational model performances from word-based datasets to ImSim-1498.

5.9 Distribution of Spearman's rank correlations between participant and ImSim-1498 ratings

6.1 Retroactive Interference (RI): The disruptive effect of new learning on recall of old learning. Proactive Interference (PI): The disruptive effect of old memory on new learning.

6.2 Examples of stimuli used in the study.

6.3 Gallery Blast task workflow.

6.4 Representation of a single learning trial in block A, a single learning trial in block B, and a single recall trial in Mezurio Gallery Blast.

6.5 Distributions of the demographic-level variables age, biological sex, education level, family history, and the SCD score.

6.6 Distributions of the stimulus-level variables object pair similarity and congruency type.

6.7 Path diagram showing the number of participants who attempted the Gallery Blast task, who were removed, and who were considered in the present analysis.

6.8 Distributions of errors made by participants on the outcome variables.

6.9 Interaction effect of age and baseline errors on susceptibility to PI and RI.

6.10 Effect of object pair similarity on susceptibility to PI and RI.

6.11 Differential progression of the effects of object pair similarity from young to old in PI and RI.

6.12 Effects of different levels of congruency on susceptibility to PI and RI.

6.13 Progressive effect of interference postulated from young and healthy to pathological aging.

List of Tables

3.1 29 semantic categories of objects used in this thesis

4.1 Summary of evaluation metrics for all computational models

5.1 Summary statistics of the reviewed datasets

5.2 Demographic characteristics of participants from GameChanger round 2 selected for analysis

5.3 Summary of correlation scores between computational models and ImageSim-1498

6.1 Demographic characteristics of participants from GameChanger round 2 selected for analysis

6.2 Variables of interest for Gallery Blast analysis

6.3 Summary results from modelling PI and RI intrusion errors

Chapter 1

Introduction

1.1 Need for new tools for early detection of AD

With a rapidly growing older population, characterisation of changes to the brain as we age is an important challenge facing neuroscientists. Pathophysiological changes in diseases such as Alzheimer's Dementia (AD) begin more than a decade before the stage of clinically detectable symptoms (Jack et al., 2010; Sperling et al., 2014). In the absence of effective drugs that can reverse the neurodegenerative changes that occur in the brain with such diseases, clinical trials at the very early 'preclinical' stages hold the most promise for the development of therapeutics to prevent or slow neurodegeneration and to allow the affected to manage their risk more effectively.

Recruitment to such clinical trials, however, is impeded due to heavy dependence on traditional biomarker screens that despite providing the most direct assessment of pathology in vivo (Jack et al., 2010; Rowe et al., 2010; Sperling et al., 2011) are costly, invasive, and not readily available (Cummings et al., 2016). In-clinic assessments using cognitive markers provide an alternative (Glymour et al., 2018) but are inconvenient and unreliable for many reasons. First, they require the presence of trained neuropsychologists, which overwhelms the healthcare system as the number of patients increases. This results in delayed and infrequent assessments that do not provide a clear picture of the state of memory and cognition. Second, such assessments are usually pen and paper scoring general-purpose measurements lacking sensitivity and are only suitable for late-disease stages. Last but not least, such visits are inconvenient and stressful for many older patients requiring travel from their homes, sometimes over long distances, many of whom face mobility challenges due to old age. This added discomfort also prompts biased responses from patients during assessment, thus leading to inaccurate results (Wild et al., 2016).

Such drawbacks of traditional techniques, therefore, warrant an urgent need for non-invasive cognitive tools that are sensitive to earliest changes in cognition and can be deployed remotely in individuals' lives over longer durations for frequent and repeated measurements. Behavioural tasks targeting cognitive domains that are at risk of earliest alterations per the underlying disease pathology are key to providing disease-specific sensitive measurements. When deployed on widely accessible digital devices such as smartphones, remote assessments of such tasks can be facilitated with ease. Digital tools can, therefore, serve as platforms to such tasks to facilitate high-frequency cognitive assessment (Doraiswamy et al., 2018; Harvey et al., 2017; Laske et al., 2015).

1.2 Earliest cortical site in AD pathology

As noted, a key requirement in designing behavioural tasks aimed at early detection of AD is targeting cognitive domains prone to earliest changes in AD pathology. As per the staging of Braak pathology, the PRc or the perirhinal cortex is the earliest cortical site of phosphorylated neurofibrillary tau accumulation in AD (Braak and Braak, 1991) followed by the entorhinal cortex (ERc) and the hippocampal regions of the anterior medial temporal lobe. This theory has received support from recent neuroimaging studies demonstrating a decrease in cerebral blood volume (Khan et al., 2015) and early structural alterations in the anterior temporal lobe in preclinical and early AD (Khan et al., 2015; Krumm et al., 2016).

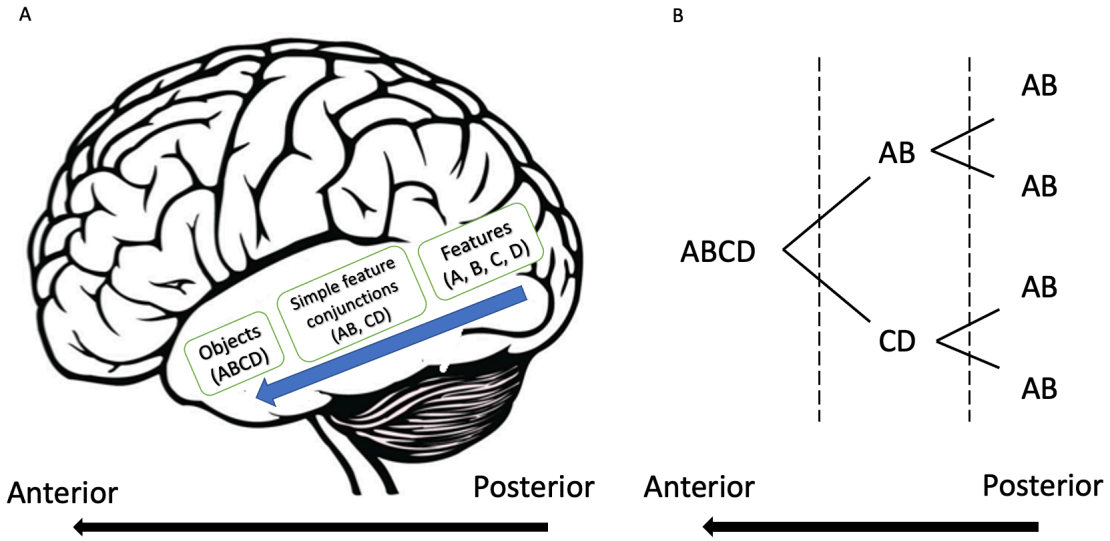


Figure 1.1. *(A) Lateral view of human cerebral cortex demonstrating the ventral visual stream showing object processing pathway according to representational-hierarchical theory. (B) Representations become more complex in more anterior regions (e.g., PRc). (Adapted from Barense et al., 2012).*

Functionally, PRc occupies the apex of ventral occipito-temporal visual processing stream. Representational-hierarchical accounts of object representation (Tyler et al., 2004; Cowell et al., 2006; McTighe et al., 2010) posit that increasingly more complex object features are coded from posterior to anterior ventral and anteromedial temporal sites (as shown in Figure 1.1), and hence the PRc is thought to contain representations of complex feature combinations needed to discriminate between confusable objects sharing a large proportion of perceptual and semantic features (Ungerleider and Mishkin, 1982; Moss et al., 2005; Saksida and Bussey, 2010; Tyler et al., 2013; Kivisaari et al., 2013). Recent work (Martin et al., 2018; Douglas et al., 2019) has also shown that PRc is the centre for integrative coding of visual and conceptual object features. In line with the PRc being the first site of AD-related changes in the brain, tasks loading on its ability to support the identification and disambiguation of similar objects in semantic memory should provide the first signal of preclinical AD (Kivisaari et al., 2013; Mortamais et al., 2016; Lancaster et al., 2020).

1.3 Cognitive tasks targeting PRc function

Cognitive tasks aimed at targeting PRc function of identification and disambiguation of similar objects are known by various names such as object discrimination tasks, mnemonic discrimination tasks or lure discrimination tasks. Such tasks leverage on the theory that complex object representations available in the PRc are needed to disambiguate between similar objects that share many perceptual and semantic features (e.g. *tiger* and *lion* share visual features such as <*has four legs*>, <*has eyes*>, <*has sharp teeth*> and semantic features such as <*is predator*>, <*roars*>, <*lives in jungle*>, etc.) and that mere availability of simpler feature combinations coded in posterior parts of the ventral object processing stream (see Figure 1.1) is not sufficient to recognise the novelty of confusable distractor stimuli. In support of this theory, studies have shown that non-availability of such object features due to atrophy of the PRc leads to an increased propensity to judge novel lures or distractors as familiar (Bussey et al., 2005; Cowell et al., 2006; McTighe et al., 2010; Kivisaari et al., 2013).

Cowell et al. (2006, 2010), for instance, have shown that loading on the ability of the PRc to distinguish between objects sharing perceptual and semantic features forms the basis of delayed matching-to-sample (DMS) recognition tasks in which stimuli shown at delay share many features with objects shown earlier (e.g., *tiger* shown first, and a similar stimulus *lion* shown at delay). Bussey et al. (2002) and Bartko et al. (2007) also found that following PRc degeneration, recognition memory performance on confusable stimuli was impaired due to non-availability of complex object representations needed for the identification of novel lure objects as different from similar stimuli encountered before. Kivisaari et al. (2013) tested this theory in healthy participants as well as in patients with varying degree of PRc damage using a recognition memory task with confusable and less confusable realistic object pictures. The study was based on the premise that living objects are inherently more confusable

than non-living objects since they share more perceptual features and therefore animacy should be used as a proxy for confusability. The authors found that while the healthy participants performed comparably in both conditions, the PRc-atrophied group committed more false positive errors with the more confusable living objects (sharing more perceptual and semantic features, e.g., *apple* and *pear* among FRUITS and *crow* and *magpie* among BIRDS) than with the less confusable non-living distractors (sharing fewer perceptual features). It was confirmed through voxel-based morphometry analysis that this effect was indeed associated with atrophy of the PRc. This study also established that AD patients exhibit a heightened propensity to commit false positive responses with inherently confusable stimuli.

In sum, since AD pathology suggests that PRc is the earliest cortical site of phosphorylated neurofibrillary tau accumulation, early AD patients (including preclinical AD patients) would perform poorer than non-AD patients on tasks that require disambiguation of confusable distractor objects with many shared features. Loading on the disease-specific trait, object-discrimination tasks use performance of individuals on their ability to disambiguate between confusable objects as a marker for early detection of AD. Next, with the help of a popular smartphone-based object discrimination task, I show an example of how these confusable objects are used in episodic memory assessment.

1.4 Case study: Mezurio Gallery Game

Mezurio (https://mezur.io/) is a smartphone application developed by the Digital Phenotyping group at University of Oxford, consisting of novel cognition measurement tasks aimed at targeting cognitive processes first vulnerable to change in AD. Among such tasks is the Gallery Game that targets the processes of episodic memory first vulnerable to neurofibrillary tau-related degeneration in AD, thus providing a sensitive

marker of PRc dysfunction (Lancaster et al., 2020). The task assesses the performance of its participants on their ability to remember learned object images over long time periods (lasting days) as well as their ability to distinguish these from confusable matched but novel distractor object images. The task design, the confusability generation scheme employed, and the findings are discussed next.

1.4.1 Task Design

Gallery Game follows a delayed recognition design. A generic delayed recognition task consists of an encoding or learning phase and a recognition or recall phase. In the encoding phase, the participants are required to learn a set of stimuli presented to them one at a time. This is followed by a delay, after which a recognition or recall task is presented in which repeat or matched but novel distractor stimuli are presented to assess whether participants can discriminate confusable distractors from the previously presented target stimuli after a delay.

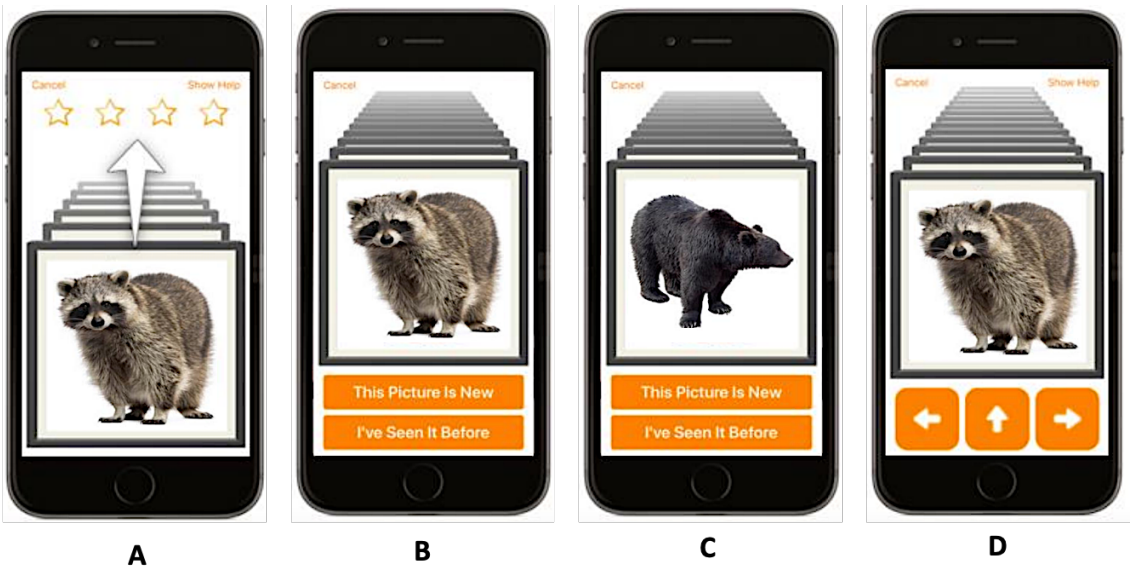


Figure 1.2. *Representation of (A) learning, (B, C) recognition and (D) recall trials in Gallery Game (courtesy: Lancaster et al., 2020).*

Gallery Game task design is unique in that it addresses the criticism that traditional in-clinic memory assessments face over shorter delay intervals (usually 30-minutes or less) between the encoding and the recognition or recall of the presented object images. The task (employed by Lancaster et al., 2020) assesses longer retention intervals and is based on the hypothesis that forgetting over longer delays (exceeding a week) could be a valuable marker for separating early AD from healthy aging. The task has a repeated-measures design following which participants are asked to complete daily assessments consisting of multiple learning tasks, each associated with a recognition and a recall phase. As shown in Figure 1.2A, the encoding phase is a paired-associate learning task in which participants are asked to learn a cued direction associated with each presented object. Upon successful learning and a predetermined delay (ranging from 10 minutes to 10 days), the recognition and recall phases associated with the object are presented. As shown in Figures 1.2B and 1.2C, in the recognition phase, participants are asked whether they have seen the presented image before. The image could be a target image learnt in the encoding phase (Figure 1.2C) or a distractor image sharing perceptual and semantic features with a target image learned in the encoding phase (Figure 1.2B). The task also consists of a recall phase (Figure 1.2D) in which the participants are asked to retrieve the swipe direction learnt during the encoding phase.

1.4.2 Confusability generation scheme

Lancaster et al. (2020) followed the confusability generation scheme used by Kivisaari et al. (2013) presented in section 1.3. The study used a binary metric of confusability based on animacy under the hypothesis that since living objects have many shared perceptual and conceptual features (e.g., <*has eyes*>, <*has four legs*>, etc.) and fewer distinctive features (e.g., <*has hump*>), they are more confusable "as a group" compared to non-living objects that have fewer features in total, with fewer shared

features (e.g., <*made of metal*>) and more distinctive features (e.g., <*is serrated*>) (McRae et al., 2005; Capitani et al., 2003; Rogers and Plaut, 2002). They matched unique target-distractor object pairs from the same semantic categories (e.g., MAMMALS, BIRDS, TOOLS, VEHICLES, etc.) as closely as possible on two psycholinguistic variables, concept familiarity and image visual complexity, to control for the influence of these variables on participant performance in object recognition (Snodgrass and Vanderwart, 1980; Kivisaari et al., 2012, 2013; Montefinese et al., 2018). However, no metric for object pair similarity was obtained. One example of confusable pairs from each domain used in the study are shown in Figure 1.3.

Figure 1.3. *Two examples of target-distractor pairs in Gallery Game (courtesy: Lancaster et al., 2020).*

1.4.3 Findings

Lancaster et al. (2020) presented an assessment of 35 healthy adults with no diagnosis of cognitive impairment (aged 40-59 years) using the Gallery Game task. The study utilised a 30-day testing phase during which participants were tested for stimulus retention intervals varying from 10 minutes to 10 days. The study found that the task was able to record behavioural differences in adults with no cognitive impairments. The study also found that binary metric of confusability (animacy) employed in the task had a significant influence on participant performance in the anticipated direction. Specifically, living distractors were indeed associated with a higher proportion of errors (i.e., were found to be harder to reject as novel and created

more confusability among participants) compared to non-living distractors. The study, however, did not comment on stimulus-level effects on participant performance.

1.5 Need for fine-grained assessment of confusability

The findings from Lancaster et al. (2020) provide support for longer-term measurements of episodic memory using Mezurio Gallery Game for detecting signals of cognitive differences associated with early AD. Additionally, the study further corroborates the use of animacy as a measure of confusability in memory assessment tasks. This reliance on a binary proxy of object confusability as opposed to a fine-grained measure of confusability between objects is a common trend seen in a number of other object recognition studies even before Lancaster et al. (2020) (Brainerd et al., 2008; Kivisaari et al., 2013; Montefinese et al., 2013; Montefinese et al., 2015). A growing body of literature, however, recognises the need for fine-grained assessment of confusability in order to facilitate a rigorous analysis of stimuli-level effects in such tasks (see Hunsaker and Kesner, 2013; Deuker et al., 2014; Liu et al., 2016; Stark et al., 2019 for support). Such an analysis is deemed essential to investigate the specific contribution of stimulus pair similarity in false recognition of confusable objects (Montefinese et al., 2017). Additionally, controlled parametric variation of confusability using precise stimulus pair similarity metrics has been shown to yield a corresponding graded behavioural response from participants in such tasks (see Chapter 2 for a review), where 'parametric' variation refers to variation controlled using a parameter, such as *time*, *distance or human-annotated similarity* between the target and the distractor. This detailed response allows for stronger conclusions to be made about cognitive differences among participants within a shorter timeframe (Liu et al., 2016; Baumann et al., 2019), which is important for solving key challenges associated with early detection of AD such as clear disambiguation between healthy aging and preclinical dementia in older adults (Sperling et al., 2010; Toner et al., 2011).

Statistically too, continuous metrics have a number of advantages over discrete (e.g., binary) variables such as more fidelity and need for fewer samples to draw inferences from, making such a setting desirable in clinical experiments.

The concept of using a continuous metric of stimulus pair similarity in mnemonic discrimination tasks is not new. In the last decade or so, a number of tasks targeting hippocampal function of lure discrimination have used scene (Bonnici et al., 2012a; Bonnici et al., 2012b; Chadwick et al., 2014) and abstract image stimuli (Holden et al., 2012; Paleja et al., 2014; Sheppard et al., 2016), varying on a parametric scale of confusability (see Chapter 2 for a review), to capture cognitive differences between healthy and at-risk participants. However, a lack of validated image datasets of confusable object pairs with fine-grained scores of object pair similarity, where the pairs are composed of similar but distinct concepts (e.g., <*lion, tiger*>), has restricted studies from using such stimuli in a continuous metric setting (Lacy et al., 2011; Montefinese et al., 2015, 2017) despite the significance of such stimuli in triggering PRc function and consequently in early detection of AD. An urgent need for carefully created and validated datasets of parametrically varying confusable object pairs is therefore warranted. Before discussing the methods to measure similarity between object pairs to create such datasets, I summarise key characteristics needed to make such datasets suitable for use in memory assessment tasks.

1.6 Characteristics of similar pair datasets

To be usable in object discrimination tasks targeting the PRc function (such as Mezurio Gallery Game), datasets of confusable stimulus pairs should follow the following characteristics:

1. ***Visual stimuli of everyday objects should be used to construct such databases***

   The PRc occupies the functional apex of the ventral occipito-temporal visual processing stream and has been shown to play an essential role in visual object recognition memory (Zola-Morgan et al., 1989; Meunier et al., 1993) and multimodal object memory (Kivisaari, Probst and Taylor, 2013). fMRI studies have shown that PRc activity is associated with fine-grained analyses of visual objects (Tyler et al., 2004), recognition of ambiguous visual objects (Moss et al., 2005) and discrimination of highly similar visual objects (Taylor et al., 2009, 2011). Studies have therefore concluded that to engage the PRc in object discrimination tasks aimed at assessing the ability to discriminate between similar stimuli, images of everyday objects are the ideal choice (Bussey and Saksida, 2002; Bussey et al., 2005).

2. ***Concepts in a pair should belong to the same superordinate category***

   Superordinate categories are high-level categories that subsume basic level concepts that share a number of features but are also distinct enough to have fundamental differences. Some examples of superordinate categories are: MUSICAL INSTRUMENTS, FRUITS, TOOLS, CLOTHING, FISH AND BIRD, while concepts such as *guitar, piano, apple, saw, screwdriver, eagle, sparrow,* and *dog* are considered basic-level concepts (Berlin, 1972; Rosch et al., 1976). A large body of literature on the working of semantic memory agrees that overlap and sharing of semantic features across modalities (visual, conceptual, and categorical) have an influence on recognition and disambiguation performance (Rogers and Plaut, 2002; Capitani et al., 2003; McRae et al., 2005; Montefinese et al., 2015). In addition to the property that members belonging to the same superordinate category share perceptual and semantic features, Loewenstein et al. (2003, 2004) have shown that persons with

preclinical AD are susceptible to interference from competing representations from stimuli belonging to same superordinate category. Therefore, a first necessity of highly confusable pairs of objects to be used in cognition assessment tasks is that the concepts in a pair should belong to the same superordinate category.

3. ***Concepts in a pair should be similar rather than merely related***

   The difference between association and similarity is important for acquiring confusable concept pairs. Pairs such as <*petrol, car*> are related or associated since they occur in similar context, but they are not similar, since they do not share visual, conceptual, or semantic features (*car* is a VEHICLE, *petrol* is a FUEL). However, the pair <*car, bus*> can be considered similar since the constituent concepts share multiple features. Preliminary data from mnemonic discrimination studies (Cann et al., 2011; Montefinese et al., 2015; Coane et al., 2016) have shown that the errors on object discrimination tasks can't be explained by association alone. Coane et al. (2016) showed that more errors were made on discriminating new stimuli from old when the confusable concept pairs shared visual and semantic features as well as categories with the lures in addition to being associated. These evidences show that to produce robust confusable pairs for use in memory assessment tasks, concepts in confusable pair should overlap on multiple facets of similarity – visual, conceptual, and semantic.

4. ***One target image should be matched with only one distractor image and vice-versa***

   An ideal dataset should have one-to-one mapping between target and distractor stimuli. This means that if a target (say *lion*) has been paired with a distractor (say *tiger*), then neither the target nor the distractor in the database can be paired with other stimuli. This is important given the setup of object

discrimination tasks in which assessment is based on whether having seen a particular target stimulus before, its matched distractor stimulus can be identified as new. Moreover, this setup is necessary for fine-grained analysis such as relationship between pair confusability and errors made on disambiguating concepts from such pairs.

5. ***Degree of similarity between stimulus pairs in a dataset should vary parametrically***

   Liu et al. (2016) assert that to generate a gradient of interference produced by the stimulus pairs used in an object discrimination task, the degree of similarity between concept pairs in a dataset should vary parametrically. This would also allow in making stronger conclusions about the relationship between degree of interference and stimulus pair similarity, than if results were solely based on a binary response such as that seen in the Gallery Game task in section 1.4.

As noted in section 1.5, while object discrimination tasks presented earlier in this chapter have covered characteristics 1 to 4 in their stimuli selection schemes to some extent, they have not addressed trait 5. Addressing this trait requires validated methods of measuring feature overlap between object pairs, only upon operationalizing which can object pairs with varied degrees of similarity be chosen and desirable datasets of confusable object pairs be generated. Next, I present an overview of the most popular method used to obtain similarity between distinct object pairs: feature norms.

1.7 Feature norms for measuring object pair similarity

Feature norms are hand-coded lists of visual and semantic features of everyday objects. Feature norms-based studies (Kremer and Baroni, 2011; McRae et al., 2005; Vinson and Vigliocco, 2008) require human participants to list as many attributes of a

given object as they can in limited time. For instance, attributes for the concept *dog* in such lists would include <*has a tail*>, <*has four legs*>, <*has teeth*>, <*barks*>, etc. Here, 'limited time' would be an experimental decision made to ensure the completion of the data collection exercise in time, as well as to limit the cost incurred (e.g., see Battig and Montague (1969), where they limited the participant input time to 30 seconds). Object feature lists from such studies are thought to represent salient psychological aspects of object meaning and are widely used as proxy for perceived sensory and behavioural properties of objects (Silberer, 2015). A number of studies have used these feature norms to derive similarity between concept pairs by measuring the overlap between concept feature vectors using metrics such as cosine similarity (Clarke and Tyler, 2014, 2015; Montefinese et al., 2017, 2018; Buchanan et al., 2019). Such a method would yield similarity values between concept pairs on a scale of 0 (minimum similarity) to 1 (maximum similarity), which could then be used to create the required database of parametrically varying confusable pairs. However, the feature-norms method has its limitations.

First, available published norms (e.g., McRae et al., 2005; Devereux et al., 2014) are based on responses to verbally presented stimuli, and currently there is no normative concept feature data gathered using object images. This makes these norms unsuitable for visual stimuli-based object discrimination studies since participants list object features in such studies by creating mental models of each concept based on their experience (Barsalou, 2003) leading to features that may be generically applicable to the concept but are not focussed on the exemplar image of a concept chosen for the object discrimination study. Second, elicitation from human subjects limits the scope of this method to a small number of concepts for which feature norms are acquired. Buchanan and colleagues (2019) suggest that human elicitation is limited to salient features and is unlikely to represent an exact representation of concept. As such, when attributes are acquired for uncommon concepts such as

different kinds of fish or similar looking tools, the subtle differences are hard to elicit for human participants in linguistic form (McRae et al., 2005).

1.8 Motivation and aim of present research

The drawbacks of feature norms-based methods warrant the need for exploration of alternative methods of obtaining object pair similarity. Such new methods would need to detect subtle differences between concepts by mapping out detailed object features that are image as well as concept-specific and are readily obtainable for previously unseen concepts. Once such features are extracted and similarities between object pairs measured, the required databases of parametrically varying confusable pairs can be constructed.

In the last decade or so, advancement in the field of artificial neural networks and deep learning has led to the production of novel computational methods that can be used to automatically generate feature vectors for a large number of concepts. While computer-vision based methods, in which connectionist models are trained on a large set of images, are being used to generate visual feature vectors for previously unseen images (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2016), techniques such as the skip-gram approach (Mikolov et al., 2013) are being used in conjunction with the distributional hypothesis (Harris, 1954) to generate linguistic feature vectors for concepts by making use of patterns of word co-occurrence in corpora to provide a proxy for semantic and relational features of concepts.

In their study on human-rated similarity between object pairs, Jozwik et al. (2017) showed that vision-based deep convolutional neural network (CNN) models are not only capable of automatically generating visual features for previously unseen objects, but also perform comparably to hand-coded feature norms when approximating human

judgement of concept pair similarity. On the other hand, owing to their considerable success in simulating human behaviour, corpus-based language models have been used for stimuli selection in word-based synonym selection and semantic priming tasks (Padó and Lapata, 2007; Bullinaria and Levy, 2012). A number of studies (Bruni et al., 2014; Silberer and Lapata, 2014; Kiela and Bottou, 2014; Lazaridou and Baroni, 2015) have also explored whether combining information from such models improves their effectiveness at estimating human behaviour over individual models of vision and language.

The utility of such models in obtaining databases of confusable object image pairs that share perceptual and semantic features, however, remains unexplored. Owing to the need for such databases for fine-grained assessment of confusability in memory assessment tasks, the present thesis explores **the potential of computational models of vision, semantic association, and taxonomy, as well as their combinations, in generating datasets of parametrically varying confusable object image pairs usable in memory assessment tasks and examines the utility of these computationally matched object pairs in an objective test of short-term memory**. Next, I outline the structure of this thesis to summarize how this assessment was carried out.

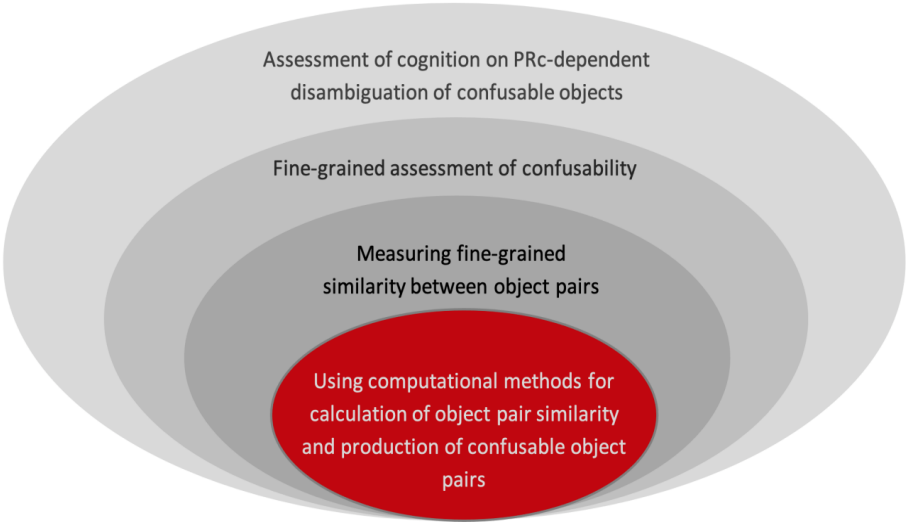


Figure 1.4. *Schematic diagram depicting scope of this thesis.*

1.9 Thesis structure

Chapter 2 presents a review of confusability generation schemes in lure discrimination tasks that use visual stimuli, including object images, natural scene, and abstract images for their assessment. Through such a review, the need, impact, and benefits of fine-grained assessment of confusability are highlighted and it is proposed that confusable object stimuli-based studies central to this thesis should also adopt this style of assessment.

Chapter 3 presents a background of the computational models of vision and semantics used in this thesis to produce datasets of confusable object pairs as well as the standard techniques used in prior work to evaluate the performance of the employed models in similar use cases.

Chapter 4 demonstrates how the models presented in Chapter 3, as well as their weighted combinations, were used to produce datasets of confusable object pairs parametrically varying in similarity in this thesis. The presented models and the resulting object pairs are also evaluated against standard techniques discussed in the previous chapter. With the help of this evaluation, a comparison among the implemented models is facilitated. The object pairs matched by the best performing models are then selected for further validation.

Chapter 5 presents a carefully designed 'human in the loop' study conducted to obtain the similarity ratings of the computationally produced object pairs from human participants. The study was hosted on the Amazon mechanical Turk platform to obtain the required similarity ratings from over 500 human participants, which were then used to further validate the efficacy of the computational models at estimating the human understanding of object pair similarity.

Chapter 6 presents an assessment on whether the matched object pairs obtained from the previous chapters inflict confusability when used in an objective cognitive test and whether increasing object pair similarity differentially impact participants from different demographics, thereby examining the possible use of this metric for sensitively recording cognitive differences in healthy adults. For this, the computationally matched confusable object pairs were deployed in a paired associate learning task based on Mezurio Gallery Game, called Gallery Blast, which was remotely attempted by a large sample of UK-based participants with varying demographics under the GameChanger study.

Chapter 7 concludes the thesis with a summary of the main findings of this thesis, and highlights avenues for further research.

Chapter 2

Background: Confusability generation schemes in memory assessment tasks

In the previous chapter, I defined the scope and the need for the research undertaken in this thesis. In the present chapter, I present a review of the confusability generation schemes employed in existing memory assessment tasks. This aim of this review was to explore the methodologies used in existing tasks for producing datasets of confusable stimulus pairs parametrically varying in similarity across stimulus types, the relationship between stimulus pair similarity and the size of interference experienced by the participants who attempted the task, and the differential effects of stimulus pair similarity experienced by the participants at different stages of cognitive health. Such a review would not only motivate the need for the work undertaken in this thesis but would also help establish the expectations from the analysis of participant performance on the memory assessment task administered and discussed later in this thesis.

This chapter is composed of four parts. In section 2.1, I present examples of confusability generation schemes employed by existing behavioural studies using the stimulus type other than that central to this thesis. The goal of this section is to demonstrate how existing tasks have reaped the benefits of fine-grained assessment of confusability for obtaining stronger conclusions of participant performance on such tasks. In section 2.2, I review the confusability generation schemes employed in tasks central to this thesis, i.e., object stimuli-based discrimination tasks designed to target the PRc function. This review highlights the drawbacks of current methods of confusability generation in such tasks and further justifies the need for the work undertaken in this thesis. In section 2.3, I present a review of studies providing possible solutions to the issues described in section 2.2. Finally, in section 2.4, I

summarise my findings including gaps in research that need to be addressed and possible ways to address them.

2.1 Fine-grained metrics of confusability in memory assessment tasks

A number of studies (Tolentino et al., 2012; Paleja and Spaniol, 2013; Reagh et al., 2014; Leal and Yassa, 2014; Pidgeon and Morcom, 2014; Roberts et al., 2014) have previously shown that by parametrically varying the input stimuli to mnemonic discrimination tasks using fine-grained metrics of confusability, corresponding behavioural response, varying from one individual to another, can be obtained. This graded response helps in sensitively recording cognitive differences among individuals attempting such tasks. Examples of the methods of parametric variation of confusability of the input stimuli employed by such studies include varying spatial location (Holden et al., 2012; Paleja et al., 2014; Sheppard et al., 2016) and degrees of rotation between the studied and lure stimuli (Motley and Kirwan, 2012), and varying the visual similarity between conceptually same input stimulus pairs (e.g., cup 1-cup 2), where the visual similarity has been recorded, for instance, through human ratings of stimulus pair similarities (Bonnici et al. 2012a; Bonnici et al., 2012b; Pidgeon and Morcom, 2014; Fujii et al., 2014). Such studies have successfully captured detailed effects of aging (Leal et al., 2014b; Leal and Yassa, 2014; Yassa et al., 2011b), as well as the differences between the performance of impaired and unimpaired individuals on mnemonic discrimination of confusable stimuli (Holden et al., 2012).

In this section, I present a review of the confusability generation schemes from the mentioned studies involving stimulus types not central to this thesis. These would include natural scene stimuli, abstract stimuli, and visually distinct but conceptually same object pair stimuli. This review will help in recording the relationships between stimulus pair similarity and the size of interference, as well as the differential

magnitudes of interference experienced by the participants at different stages of cognitive health when attempting memory assessment tasks employing the mentioned stimulus types.

2.1.1 Fine-grained assessment using natural scene stimuli

Natural scene stimuli have been used as inputs to mnemonic discrimination tasks for a number of purposes, including, but not limited to, examining the role of medial temporal lobe subfields in distinguishing between two similar scenes (Bonnici et al., 2012a; Bonnici et al., 2012b), assessing the relationship between size of hippocampal subfields and memory recall (Chadwick et al., 2014), studying match enhancement and attentional modulation in the human medial temporal lobe (Dudukovic et al., 2010), and exploring the difference between the role of the hippocampus and MTL cortex in classifying different categories of images (Huffman and Stark, 2014). In such studies, methods such as parametric morphing of scene stimuli as well as normative scoring methods of rating the similarity between pre-selected natural scene pairs by human participants have been used to generate continuous metrics of confusability.

In one such study, Bonnici et al. (2012a) inspected participant behavioural response on parametric variation of similar scene stimuli. In this study, young and healthy adults (n = 16, mean age = 24.4) were asked to choose whether a presented scene was more likely to be scene A or scene B (see Figure 1 in Bonnici et al., 2012a) and provide a confidence rating of their decision. To generate confusability, two distinct scene images (see Figure 1a in Bonnici et al., 2012a) were morphed using the computer software “Morph Age” to generate confusable images in a continuous fashion with complimentary contribution of the two scenes (10% scene A + 90% scene B, 20% scene A + 80% scene B, etc.) as shown in Figure 1 in the article. The idea behind the behavioural analysis was to inspect whether with progressive difficulty,

participants perform continuously worse on reaction time and the confidence in their choice.

The overlap between the two scenes in each image was used to determine the confusability metric of the morph. For instance, 100% scene A and 100% scene B were the least confusable stimuli, with confusability increasing with overlapping scenes such that 50% scene A + 50% scene B was the most confusable image. The authors reported that the interference due to confusability followed the hypothesised pattern, i.e., with an increase in confusability, there was a corresponding decrease in participants' confidence in their choice and an increase in their reaction times. Parametric control over confusability thus helped the authors obtain precise functions of relationship between confusability and interference in memory (see Figure 2 in Bonnici et al., 2012a), thus allowing them in reaching strong conclusions about behavioural response of young and healthy participants on processing of similar scene stimuli.

In another study, Chadwick et al. (2014) showed confusable pairs of video clip stimuli to young and healthy participants (n=15, mean age=21.17) and asked them to recall the videos during an fMRI scan. The purpose of the study was to assess the relationship between the sizes of hippocampal subregions and participant performance on disambiguation of similar stimuli. The stimuli video clips had overlapping foreground and background scenes, with each pair rated for similarity beforehand by human participants. The authors found that a continuous variation in stimulus pair similarity helped them discover that participant performance on the task varied proportionally to the size of the hippocampal region of interest.

In a similar vein, Leal et al. (2014b) used normative human ratings of stimuli to find the effects of aging on mnemonic discrimination of emotional information. The authors used a set of scene stimuli, rated beforehand by a normative group on emotional valence on a Likert scale of 1-9 (1 being negative, 5 being neutral and 9 being positive). Different cohorts of young (n = 38, mean age = 21) and healthy older adults (n = 38, mean age = 68.34) were asked to learn a set of stimuli in the encoding phase. In the test phase, the confusable counterparts were mixed with the original stimuli to assess memory function and difference in participant response on emotional valence. The authors found significant effects of age in discrimination of confusable objects, but interestingly among older adults, they found that increasing emotional valence (both positive and negative) increased their discrimination capacity compared to neutral stimuli in a continuous fashion. Fine-grained scales of measurement, therefore, helped the authors determine the specific differences among young and old cohorts in their processing of emotional stimuli as well as in obtaining differences in their forgetting curves.

In sum, studies using natural scene stimuli parametrically varying in confusability discussed in this section obtained detailed curves of participant response, as well as precise differences between participants belonging to different cognitive and age groups.

2.1.2 Fine-grained assessment using abstract image stimuli

Abstract images are defined as shapes, silhouettes, or lines not easily identifiable as everyday objects or scenes (Liu et al., 2016). Abstract stimuli have been used in mnemonic discrimination tasks to examine the effects of aging on implicit and explicit memory for novel visual objects (Schacter et al., 1992), to find evidence for the role of dietary interventions in enhancing cognitive function in older adults (Brickman et al.,

2014), to examine the difference between information representation in the hippocampus and the MTL cortex (Duff et al., 2012), to clarify which hippocampal subregions are functionally associated with depressive moods in humans (Fujii et al., 2014), and to assess age-related changes to varying degrees of spatial interference (Holden et al., 2012; Paleja et al., 2014; Sheppard et al., 2016), temporal interference (Tolentino et al., 2012) and semantic interference (Pidgeon and Morcom, 2014) in behavioural discrimination tasks.

Among the confusability generation schemes using abstract stimuli, spatially varying the location of a stimulus is a popular method of parametrically controlling confusability on a continuous scale in a behavioural discrimination task. Typically, in the learning phase, participants are shown an abstract shape such as a circle, a square or a dot on a screen and are expected to memorise its location. In the test phase, the same abstract shape is presented in either the original or a distant location, and the participants are expected to indicate whether the shape is in the original or a distant location. The distance of separation between the target and the distractor locations controls the level of confusability (confusability decreasing with distance of separation from the target location).

Holden et al. (2012), Paleja et al. (2014) and Sheppard et al. (2016) have used this method to parametrically vary confusability in their recognition experiments. Holden et al. (2012) used the spatial location task to assess differences in spatial pattern separation in young and older adults. The four possible spatial separation distances (or parameters) employed in their experiment at test were 0 cm, 0.5 cm, 1.0 cm, and 1.5 cm. The authors found a corresponding relationship between decreasing interference (or increasing distance) and increasing recognition accuracy across both age groups. The authors further found that using this parametrically varying setup,

they were able to obtain clear differences between behavioural responses of younger (n = 30, mean age = 20.1 years) and older adults (n = 30, mean age = 74.5 years). Moreover, upon further analysis, in which older adults were classified as healthy and cognitively impaired using a standard test of cognition, the authors found that older unimpaired participants did not significantly differ from younger participants in higher interference conditions but only in lower interference conditions (distance of 1cm and above) as shown in figure 2 in the article. These relatively continuous curves of performance on the task show that somewhere between 0.5cm and 1cm, differences in cognition between older unimpaired and younger participants begin to manifest. These results show how parametric variation of confusability using abstract stimuli has helped researchers capture group-level differences among participants on such tasks.

Sheppard et al. (2016) used the same setup as Holden et al. (2012) in their task to assess spatial pattern separation differences in healthy young as well as older impaired and unimpaired APOE-ε4 carriers and non-carriers. As with Holden et al. (2012), they too found that accuracy on the test increased as a function of decreasing interference across all groups. Additionally, they found that older impaired ε4 carriers performed significantly worse than all other groups across degrees of confusability.

Paleja et al. (2014) used a similar setup with one key difference – they used a binary measure of interference, near (2 distance units apart) and far (4 distance units apart), instead of more levels of interference used by the other two studies. While Paleja et al. (2014) too found that accuracy for participants in recognising the right locations increased with a decrease in confusability (near: high confusability, far: low confusability), unlike other studies, they were unable to obtain more detailed differences in participant response.

In another method of parametrically varying confusability between abstract stimulus pairs, Pidgeon and Morcom (2014) and Fujii et al. (2014) used line drawings of everyday objects and monochrome silhouette figures respectively in their behavioural separation tasks to assess the impact of confusability on interference in memory. Both these studies used subjective human participant ratings for determining the similarity between confusable stimulus pairs. Pidgeon and Morcom (2014) assessed the effects of aging in false recognition of perceptually and conceptually similar images. Perceptual and conceptual similarities between confusable object pairs were obtained on a 5-point Likert scale. The authors found that this helped them obtain a detailed assessment of relationship between stimulus pair similarity and false recognition. The authors also found clear differences in cognitive responses between young and older adults across fine-grained similarity values, i.e., older adults performed worse than young adults with increasing similarity between stimulus pairs.

In a similar vein, Fujii et al. (2014) used monochrome silhouette figures in their behavioural discrimination task to assess mnemonic discrimination among young and cognitively intact individuals (n = 27, mean age = 22.52) in depressive moods. Using a typical delayed matching-to-sample task, participants were expected to learn an image at encoding and retrieve it at test. Image pairs were rated for similarity by an independent group of young adults on visual analogue scale (VAS) of 10 cm length. The authors found a corresponding relationship between stimulus pair similarity and interference produced by such pairs.

In sum, as observed with scene stimuli, detailed similarity ratings between abstract image pairs too had corresponding effects on interference in memory across different age and cognitive groups. Studies using this stimulus type, where inspected, found a) a direct relationship between stimulus pair similarity and the magnitude of interference

they produced and b) that the older impaired performed worse than older unimpaired, who in turn performed worse than the young cohort on the tasks employing this stimulus type. In addition, such continuous scales helped obtain detailed forgetting curves, thereby helping scholars find precise differences between participants belonging to different cognitive and age groups.

2.1.3 Fine-grained assessment using everyday object pairs (related at basic category level)

Based on category membership of object pairs, there are two variants of object discrimination tasks: 1) tasks in which the studied and lure objects are related at basic category level and differ only in visual features (e.g., two *cats* differing in visual feature such as colour); and 2) tasks in which the studied and lure objects are related at superordinate category level and differ both visually and conceptually (e.g., *lion* and *tiger*). First, I discuss tasks belonging to the former category along with a review of their confusability generation schemes. A review of the latter is presented in section 2.2.

Everyday object stimulus pairs related at basic category level have been used in mnemonic discrimination tasks to assess the role of hippocampal subregions in different processes in the brain (Kirwan and Stark, 2007; Bakker et al., 2012; Duncan et al., 2012; Motley and Kirwan, 2012; Kim and Yassa, 2013; LaRocque et al., 2013; Manelis et al., 2013; Staresina et al., 2013; Huffman and Stark, 2014; Reagh and Yassa, 2014; Doxey et al., 2015; Bennet et al., 2015; Bennet et al., 2016), in studying the role of age-related (Toner et al., 2009; Yassa et al., 2011a; Yassa et al., 2011b; Holden et al., 2013; Pidgeon and Morcom, 2014; Bowman and Dennis, 2015; Reagh et al., 2014; Roberts et al., 2014; Stark et al., 2013; Stark et al., 2015) and neurological disorder-related (Kivisaari et al., 2012; Kirwan et al., 2012; Kivisaari et al., 2013;

Shelton and Kirwan, 2013; Das et al., 2014; Reagh et al., 2014; South et al., 2015; Martinelli et al., 2015) reductions in human recognition memory, effects of dietary supplements on cognition (Segal et al., 2012; Borota et al., 2014) and validating the task paradigm itself in various settings (Reagh et al., 2014). Studies using this stimulus type have relied on both objective and subjectively obtained parameters to measure and vary stimulus pair similarity. Objective methods include spatial and rotational variation in stimuli, whereas subjective rating and false alarm-based methods have been used in producing normative ratings of similarity between stimulus pairs.

Motley and Kirwan (2012) used rotation as a parameter to vary the confusability between image pairs of everyday objects in their mnemonic discrimination experiment. Their aim was to define the response function to parametrically varying target-lure similarity in healthy, young adults. Target images were rotated by 15, 25, 35 and 55 degrees to parametrically create distractor images, with images rotated by 15 degrees being the most similar to the target image and hence the most confusable, and the ones rotated by 55 degrees being the least similar and hence the least confusable. A five-scale measure of confusability (0, 15, 25, 35, 55 degrees) helped the authors not only in creating a corresponding behavioural response function, but also in quantifying the thresholds of hippocampal sensitivity to stimulus change. The authors found that aligning with the expectation, accuracy was positively correlated with increasing angle of rotation. In other words, even with this stimulus type, confusability was inversely proportional to participant accuracy on target-distractor disambiguation.

In another study, Reagh et al., (2014) used varying spatial distance between the first and subsequent presentation of the stimuli in a delayed recognition task to induce confusability. The aim of their study was to characterise detailed differences among young, older unimpaired and older impaired adults on their mnemonic discrimination ability. This setup is similar to the spatial separation method reported in section 2.1.2

for evoking confusability in abstract stimuli. The authors divided the screen into 35 equal and imaginary grids. As shown in figure 1 in the article, the image could be present in separate grids during encoding and retrieval. The participants were expected to indicate whether the distractor was in the same location as the target stimulus. The number of grids that the distractor was away from the target (the distance metric) was used to measure confusability – fewer the number of grids between the target and the distractor, more the confusability. As with other studies, the authors reported positive relationship between degree of confusability and degree of interference produced by such pairs. Also, the parametric variation of confusability helped the authors map-out the specific differences among the three cohorts as they found a performance deficit between young and healthy older adults only beyond 2 units of separation between target and distractor stimuli, while impaired older adults performed worse than both the other groups across all levels of confusability (figure 4 in Reagh et al., 2014).

Another common method for obtaining “true” confusability produced by object pairs is acquiring baseline interference produced by pairs of stimuli with a normative, young group. For this, a set of young and healthy human participants are recruited to perform the intended behavioural experiment and then stimulus-wise false alarms produced in a recognition test are used to produce the ideal metric of similarity between stimuli in a pair. Participants from at-risk population can then be assessed on how they perform in comparison to the healthy cohort.

Yassa et al. (2011b), for instance, recruited a group of college-age undergraduates to participate in a delayed recognition task to generate a baseline mnemonic confusability rating for each pair. The more errors made on a stimulus pair by the young, healthy cohort, the more confusable it was rated and vice versa. The authors then used the

same set of stimuli in older cognitively unimpaired participants to understand the effects of aging on object discrimination capabilities. Finer assessment helped the authors show that the relationship between stimulus pair similarity and interference produced by each pair is not only direct, but also non-linear. In addition, they were able to capture finer differences between performance of younger and older adults on this task.

Finally, obtaining subjective human-rated similarities between pre-selected stimulus pairs using Likert and visual analogue scales is another common method employed by studies to vary similarity and has already been discussed in section 2.1.2. In such a similarity rating experiment, participants independent from the memory assessment study are presented stimulus pairs and asked to rate their similarity on a Likert scale. Studies such as Duncan et al. (2012) and Shelton and Kirwan (2013) used this methodology to obtain visual similarity ratings for the object pairs used in their experiments to assess the relationship between object pair similarity and interference produced by such object pairs. In agreement with other reviewed studies, here too, the authors found a direct relationship between object pair similarity and the interference produced. Shelton and Kirwan also found that this effect of similarity was greater for the impaired participants compared to the unimpaired participants.

To conclude, as observed with previous stimulus types, using everyday object pairs varying on a continuous scale of similarity too helped studies in extracting finer details of relationships between stimulus pair similarity and interference, as well as in obtaining clear patterns of differences in behavioural responses among participants belonging to different cognitive groups.

2.1.4 Key inferences

Citing influential studies employing mnemonic discrimination tasks, I have established in this section that parametric variation of confusability has helped researchers obtain stronger conclusions of participant performance on their tasks, in that it helped them not only establish the relationships between magnitudes of confusability and interference (unanimously found to be a direct relationship), but also to sensitively record behavioural differences among participants in different states of cognitive health, viz. older impaired adults worst affected by increasing confusability followed by older unimpaired adults, in turn followed by young, healthy adults. Next, I discuss confusability generation schemes used in mnemonic discrimination tasks central to this thesis, i.e., tasks targeting PRc function in which objects in a pair vary not only visually, but also conceptually.

2.2 Confusability generation with object pairs related at superordinate category level

In the previous chapter, it was discussed how object representations are stored in PRc-dependent semantic memory. As explained, models of organisation of object memory (Tyler et al., 2004; Cowell et al., 2006; McTighe et al., 2010) agree that combined visual and conceptual features (and not one or the other) form complete representations of objects in cortical regions first vulnerable to atrophy in neurodegenerative diseases (Ungerleider and Mishkin, 1982; Moss et al., 2005; Saksida and Bussey, 2010; Tyler et al., 2013; Kivisaari et al., 2013). It was also discussed how recent work using detailed fMRI techniques has posited that integrative coding of visual and conceptual object features happens in the PRc (Martin et al., 2018; Douglas et al., 2019). Lastly, models of semantic memory, too, agree that meaning overlap and sharing of features across multiple modalities have an influence on recognition and disambiguation performance (Rogers and Plaut, 2002; Capitani et

al., 2003; McRae et al., 2005; Montefinese et al., 2015). This argument is also supported by Hunsaker and Kesner (2013), who believe that human participants make use of linguistic information to mentally describe objects, thus involving language attributes in addition to sensory or perceptual attributes into discrimination performance. It is, therefore, imperative that mnemonic discrimination tasks targeting cognitive processes facilitated by brain regions first vulnerable to tau accumulation in AD employ confusable pairs that overlap and vary parametrically on both visual and conceptual features (and not one or the other) for a thorough assessment of semantic memory.

Having established the need for this stimulus type in memory assessment, in this section, I discuss confusability generation schemes employed by object discrimination studies using such stimuli in their assessment. Additionally, in order to demonstrate the current state of research in producing datasets of confusable object pairs, I also review the methodologies used by such studies to manufacture such datasets.

Everyday objects related at superordinate category level have been used to examine the role of PRc in visual object recognition (Zola-Morgan et al. 1989; Meunier et al. 1993) and discrimination between objects sharing many features with one another (Bussey and Saksida, 2002; Bussey et al., 2005), to predict atrophy of the PRc (Kivisaari et al., 2013), to assess differences between the roles of perceptual and semantic relations between object pairs in inducing confusability (Koutstaal et al., 2003; Pidgeon and Morcom, 2014; Montefinese et al., 2015; Montefinese et al., 2017; Montefinese et al., 2018; Bowman et al., 2019), to evaluate object-specific semantic encoding in human PRc (Clarke and Tyler, 2014, 2015) and to design new cognitive tasks for early detection of AD (Curiel et al., 2013; Loewenstein et al., 2016; Lancaster et al., 2020). I next present the techniques used by such studies to generate confusability and match object pairs.

2.2.1 Binary measure of confusability

As discussed in the previous chapter, a large number of studies using object image stimuli employ a binary measure of confusability. Such studies rely on the theory that the stimuli belonging to living and non-living categories place differential demands on the PRc function, thereby creating a binary metric of confusability (living objects more confusable than non-living). Task design, nevertheless, requires them to match targets and distractors on some variables so that matched distractor object images can be used in retrieval phase to test participant memory.

Kivisaari et al. (2013), for instance, paired target-distractor stimuli according to 'visual similarity' such that one-third of their distractors were from the same semantic category and were similar in form and colour (e.g., *mouse - guinea pig*), one-third of the targets were paired with a distractor from the same category and were visually dissimilar (e.g., *giraffe - hedgehog*), and one-third of the distractors were visually similar to the target, but represented objects from a different category (e.g., *telescope - cucumber*). However, such pairing was done in-house by the authors and in their analysis, they relied on animacy as the sole criterion for confusability. Furthermore, the authors did not attempt to quantify the similarity between such object pairs and hence could not investigate the specific role of target-distractor similarity in false recognition of confusable stimuli in their task. Nevertheless, the authors did find significant effects of animacy on participant performance. From the perspective of stimulus pair production on a continuous scale of confusability, however, manual matching of target-distractor pairs limits the scope and applicability of this method in creating larger datasets of confusable pairs. In a study inspired by Kivisaari et al. (2013), Lancaster et al. (2020) used confusable object stimuli in a novel object recognition task, the Mezurio Gallery Game, to test long-term memory in 35 healthy adults (aged 40–59 years) in a month-long assessment. The task is designed on the principle of longitudinal repeated measurement and hence demands that the pairs be novel to the participant each time,

resulting in need for a large dataset of matched object pairs. Given the limitations of manual methods in producing large datasets, Lancaster and colleagues used an automated method to produce their object pair dataset. Their pipeline was as follows. First, the authors acquired a large image dataset of objects (1,200 object images) across semantic categories through a manual search on Shutterstock, a popular online image platform. To match object pairs, they first collated values of three psycholinguistic variables: metrics of familiarity, visual complexity, and semantic category for each image stimulus. They obtained the first two automatically, approximating familiarity for each concept through its published word frequency using the freely available Google n-gram tool and approximating visual complexity of each image as its compressed file size, while semantic categories were coded in-house. Upon segregating the dataset of object images into semantic categories, the authors then matched target-distractor pairs as closely as possible on visual complexity and familiarity ratings available for each image using the Munkres algorithm to produce the assignment solution with the smallest difference cost (Munkres, 1957). The authors thus produced a dataset of more than 500 image pairs to be used in their memory assessment task.

While the object pairs in this dataset shared category membership, there was no measurement of the degree of similarity among such pairs of objects. The authors thus relied on animacy as a measure of confusability for their assessment and no stimulus-level analysis of confusability was performed. This is justified since the study targeted long-term memory and used retention interval (varying between 10 minutes and 10 days) between the first and subsequent presentation of a stimulus as a parameter, the authors may have deemed animacy as a sufficient measure of confusability to sensitively record behavioural differences in adults, without the need to delve into fine-grained assessment of confusability. However, were such a fine-grained metric of

object pair similarity available to the authors, they could use it for obtaining stronger conclusions of participant performance on the task at the stimulus-level.

2.2.2 Granular assessment of similarity between distinct object pairs

As discussed in the previous chapter, in the last decade or so, a growing body of literature using object stimuli has stressed on the adoption of detailed metrics of confusability in their analysis. Such studies have recognised the need for operationalising semantic similarity between objects to better understand the role of confusability in false recognition of objects (Brainerd and Reyna, 2002). A number of such studies have relied on hand-coded feature norms (Frenck-Mestre and Bueno, 1999; Vigliocco et al., 2004; Kremer and Baroni, 2011; McRae et al., 2005; Montefinese et al., 2013; Devereux et al., 2014) to quantify similarity between object pairs by measuring cosine similarity between feature lists of such concept pairs (see section 1.7 for details). Such a method yields fine-grained similarity values on a scale of 0 (minimum similarity) to 1 (maximum similarity), which are then used for detailed inspection of the effects of object pair similarity on participant response.

For instance, Montefinese et al. (2015) calculated cosine similarity between feature vectors of concept pairs used in their study to show that concept pairs with higher similarity on feature norms were harder to disambiguate compared to those with lower similarity. In another application, Montefinese et al. (2018) further used these fine-grained values of similarity between concept pairs to manipulate the degree of featural similarity between new and old concepts, which helped them precisely investigate how object pair similarity modulates pupil dilation during false recognition of confusable objects in an object discrimination task. Similarly, Clarke and Tyler (2014) used feature norms to generate concept confusability scores, which were mean cosine similarities between feature vectors of all concepts and the target concept. These scores reflected

the number of features a concept shared with all other concepts (e.g., <*is green*> between *spinach*, *kale,* and *cabbage*). Using such confusability scores, they demonstrated the differences in object disambiguation capabilities between posterior and anterior ventral cortex. However, such studies have limited their assessments to word stimuli of objects, while as established earlier in this thesis, targeting PRc function requires use of image stimuli in memory assessment tasks. As a result, investigation targeting the use of image stimuli in object discrimination tasks on fine-grained operationalisation of similarity remains unexplored.

Where image stimuli have been used and semantic similarity between images in each pair operationalised, specific effects of confusable pair similarity have not been assessed. Bowman et al. (2019), for instance conducted a subjective human study to rate confusable pairs on similarity between members of each pair, including similarities in physical appearance, use, category, and other important characteristic on a 5-point Likert scale. Like Kivisaari et al. (2013), the authors pre-selected 'thematic lures' (or confusable distractors) in-house (see figure 1 in Bowman et al., 2019 for their confusable object triplets). However, the authors used the obtained object pair similarity only to confirm that thematic lures were not rated more similar to the targets compared to item lures and did not perform a detailed analysis of object pair similarity on participant performance in old/new object disambiguation.

2.2.3 Observations

In this section, I reviewed studies using confusable object pairs related at superordinate category level. Broadly, such studies either relied on a binary metric of confusability (e.g., Kivisaari et al., 2013 and Lancaster et al., 2020) or recognised the need for fine-grained assessment of confusability and made attempts at obtaining such metrics for their object pairs (Clarke and Tyler, 2014; Montefinese et al., 2013, 2015).

Among the latter, there was a strong reliance on human-generated feature norms to measure feature overlap between object pairs as a surrogate for object pair similarity. However, as discussed in the previous chapter, this method has its limitations. For instance, I highlighted in the previous chapter that since such feature norms were collected using only verbally presented stimuli and were not specific to any given image of an object, they were only suitable for word-based tasks. This reflected in my literature search as the only studies I found (Clarke and Tyler, 2014; Montefinese et al., 2013, 2015, 2018) making use of feature norms for measuring concept pair similarity used word stimuli. Additionally, due to the limitations of human elicitation of object features (Buchanan et al., 2019; discussed in the previous chapter), feature norm databases are also limited in size. As a result, where studies wanted to produce large datasets of confusable object pairs sharing visual and semantic features (as in Lancaster et al., 2020), feature-norm based datasets were not a suitable alternative. In sum, an exploration of newer methods to produce the discussed datasets of confusable object pairs is warranted and may yield datasets capable of feeding longitudinal studies hosting the memory assessment tasks central to this thesis over an extended period of time.

It is worth mentioning at this point in this chapter, that the subjectively measured "perceptual" similarity between objects could be affected by factors such as ethnicity and other background traits of the rater. A key reason for this could be the difference in the features of the presented objects that the raters from different cultures may find more important due to their environment or upbringing. An example of this can be found in the collection of Spanish feature norms by Vivas et al. (2017), who found that when listing the features for the concept *accordion*, while the Spanish annotators listed the feature <tango> very frequently, they never mentioned the concept <polka>, which was the opposite to the feature norm collection in the English language by McRae et al. (2005), thereby showing how culture can influence how people perceive objects

and as a result the similarity between them. As a result, a number of attempts have been made to collect feature norms and similarity values in languages apart from English; Italian: Montefinese et al. (2013), German (Kremer and Baroni, 2011), Dutch (Ruts et al., 2004), and Spanish (Vivas et al., 2017).

2.3 Automated models of object representation

The need for automatically capturing mental representations of concepts has received a lot of attention in the last decade or so. As explained in the previous chapter, a number of studies have turned to computational models of vision and semantics to automatically capture object-specific representations that could closely mimic those obtained through human elicitation (such as in feature norms). Such studies claim that not only does this methodology address the shortcomings of human elicitation discussed earlier but also provides object image-specific representations – a functionality that current feature norms lack (Bruni et al., 2014; Silberer and Lapata, 2014; Kiela and Bottou, 2014; Lazaridou and Baroni, 2015; Jozwik et al., 2017).

As discussed in the previous chapter, while some of such studies are using computer-vision based convolutional neural network models to generate visual features for previously unseen images (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2016), others are using language-based distributional models (Mikolov et al., 2013) to capture semantic features of objects. To affirm the use of such models as possible replacements for human listed feature norms, Collel and Moens (2016) assessed which specific categories of human elicited object features from McRae feature norms each of these models capture. They found that while vision-based models were better at capturing *form* and *surface*, and *colour* and *motion* attributes, language-based distributional models were better at representing *encyclopaedic* and *function* attributes. Since such models of vision and language were found to extract

complementary information, the authors asserted that combining information from such sources is the ideal way to obtain more complete concept representations.

Studies using this technique of combining information from the mentioned sources of vision and semantics (Bruni et al., 2014; Silberer and Lapata, 2014; Kiela and Bottou, 2014; Lazaridou and Baroni, 2015) tend to pursue the following steps to produce concept representation. First, they source lists of concepts they need to estimate representations for. Next, they source images closely representing these concepts. Following this, they obtain image and linguistic representations for such concepts using the computer-vision and linguistic models mentioned earlier in this section. Finally, they combine such representations using techniques such as concatenation (Kiela and Bottou, 2014; Lazaridou and Baroni., 2015) and averaging (Bruni et al., 2014; Kiela and Bottou, 2014) to obtain more complete concept representations.

Representations obtained using such techniques have been used in a number of natural language applications, such as document classification (Klementiev et al., 2012) and information retrieval (Manning et al., 2008), as well as in cognitive applications such as semantic priming and synonym selection (Padó and Lapata, 2007; Bullinaria and Levy, 2012). However, even after concerted literature searches, such techniques were not found to be used in the measurement of object image pair similarity for the construction of datasets of confusable pairs useable in object discrimination tasks. Given the need for confusable object pair production identified in this thesis, assessment of the utility of such models of automated concept representations for confusable pair production is warranted.

2.4 Discussion

A review of confusability generation schemes across stimulus types suggests that parametric alteration of confusability between the study and test stimuli helps obtain detailed response function to inputs, which in turn helps studies sensitively record behavioural differences among individuals belonging to different cognitive groups. The reviewed studies showed a direct relationship between stimulus pair similarity and the magnitude of interference produced. The studies also showed that the older impaired population tends to be affected more from increasing stimulus pair similarity than the older unimpaired population, which in turn is more affected than the younger population. To infer such patterns, obtaining fine-grained measures of similarity between stimuli in each pair was found to be imperative. Studies presented in this review either obtained these values subjectively after pre-selecting stimulus pairs or employed generative methods of variation such as parametric morphing, controlled variation of spatial distance and angle of rotation or matching objects belonging to the same semantic category on psycholinguistic variables.

However, I found a clear shortage of confusable pair production schemes for object stimuli-based discrimination studies that could carefully control for feature overlap between object pairs. As summarised in section 2.2.3, such schemes were either producing confusable pairs manually, a method limited in scope and prone to biases, or relying on feature production norms, which while valid, are limited by their small size, non-availability for object image stimuli and genericness to a concept. In another application, while Lancaster et al. (2020) produced a large database of confusable object pairs, they operationalised semantic similarity between concepts in each pair simply as category membership. They relied on the assumption that since most category members share semantic features, they must be similar. However, they did not actually quantify the degree of semantic similarity among pairs of concepts.

As a result, in section 2.3, I proposed the use of automated models of object representation for the production of datasets of confusable object pairs. I introduced the utility of these models in estimating the visual and semantic aspects of object representations and raised the need to assess their efficacy in the production of datasets of object pairs sharing visual and conceptual features that can be used in memory assessment tasks such as those demonstrated in this chapter.

2.5 Conclusion

In this chapter, the reviewed studies demonstrated that parametric regulation of confusability helps in designing robust memory assessment tasks. However, a lack of validated image datasets of confusable object pairs with fine-grained scores of object pair similarity has restricted object discrimination studies from using such stimuli in a continuous metric setting. Current methods of creating such datasets suffer from a number of limitations. There is, therefore, a need to incorporate novel and automated methods to obtain fine-grained metrics of object pair similarity to produce such datasets. I propose to assess the utility of computational models of vision, semantic association, and taxonomy, as well as their combinations, for this purpose.

Chapter 3

Background: Computational resources and techniques

3.1 Introduction

In the previous chapter I proposed that computational models of vision and semantics can be used to produce datasets of perceptually similar object pairs. I briefly discussed how object representations obtained from such models can be used to quantify the similarities between pairs of objects per these models, in-turn helping compile the required datasets of object pairs parametrically varying in pairwise similarity. In this chapter, I present a background of the computational models used in this thesis to produce such object pair datasets. I also discuss the standard techniques used in prior work to evaluate the performance of the employed models in similar use cases.

This chapter is structured as follows. In section 3.2, I elaborate on the working of the standalone computational models of vision and semantics. In section 3.3, I discuss the working of the algorithm used in this thesis to create the required datasets of unique object pairs once pairwise similarities were obtained from the computational models. In section 3.4, I discuss the evaluation techniques used in existing literature to assess the effectiveness of these computational models at estimating the human understanding of object pair similarity. These techniques have been used in later chapters to facilitate a comparison among the computational models for further analysis. Finally, I conclude the chapter in section 3.5.

3.2 Computational models of object pair similarity

Improved potential of computational systems and the consequent advent of deep learning in the last decade or so has given rise to a number of models attempting to automatically construct representations of object concepts. While some of these

models make use of information from object images to approximate the visual properties of objects, other models rely on concept co-occurrence information from text corpora and categorical information from taxonomies to capture semantic relationships between objects. In this section, I present the working of three such models central to the research presented in this thesis: language-based distributional semantic models (section 3.2.1) that automatically construct semantic representations from linguistic input such as text corpora, deep convolutional neural network models (section 3.2.2) that transform images into visual feature vectors, and lexical semantic networks (section 3.2.3) that rely on categorical relationships between lexicons to define the similarity between concepts.

3.2.1 Distributional semantic models for linguistic similarity

Distributional semantic models are based on the distributional hypothesis (Harris, 1954) that postulates that words appearing in similar linguistic context (for instance, *bread* and *butter*, *apple* and *orange*) tend to have related meanings and are thus semantically related. This hypothesis is influenced by the notion that linguistic input is key to how humans acquire the knowledge of relationships between concepts (Silberer, 2015). A large body of literature has provided empirical and theoretical support for this hypothesis (see Landauer et al., 2007 for details). Upon applying this hypothesis to large text corpora using machine learning algorithms, vector space models (VSMs) can be constructed automatically (as by Turian et al., 2010; Huang et al., 2012; Mikolov et al., 2013). These can then be used to retrieve the relationships between words in the vector space using vector similarity measures as shown in Figure 3.1.

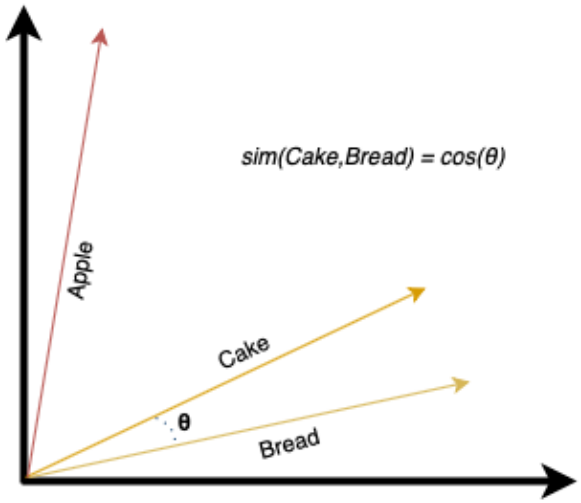


Figure 3.1. *Diagrammatic representation of linguistic vector space models in two dimensions.*

Vector spaces: A vector space is a set whose elements, referred to as *vectors* (quantities with both magnitude and direction), can be subject to mathematical operations either with a scalar or with other vectors under principles of linear space algebra. For example, given the multiplication of a vector **v** in a vector space with a scalar *a* results in a vector ***av*** with the same direction as vector **v**, but *a* times the magnitude. The addition of two vectors **v** and **w** in this vector space, however, results into a third vector **v** + **w** calculated per the triangle law of vector addition with magnitude $\sqrt{v^2 + w^2 + 2vw\ cos\theta}$ and an angle of ϕ governed by $tan\phi = \frac{wsin\theta}{v + wcos\theta}$ where

v = magnitude of vector **v,**

w = magnitude of vector **w,**

θ = angle between vectors **v** and **w**.

A vector space can be defined as finite-dimensional if its dimension is a natural number.

One machine learning technique that makes use of the distributional hypothesis is Word2Vec. This technique has been used for obtaining linguistic representations of object concepts in this thesis. Word2Vec (Mikolov et al., 2013) uses a shallow, two-

layer feedforward neural network (see Figure 3.2) to reconstruct linguistic contexts of words, with the aim of producing word embeddings (or word feature vectors) for each concept encountered in the input text corpus. The network makes use of the skip-gram model to learn the linguistic representation for each concept. In this model, the network takes as its input a large corpus of text (such as the British National Corpus). A sliding window approach is used to move through the training text, dividing the text into smaller, manageable sequences. Each word in such a sequence or "context window" is taken as an input by a softmax classifier that predicts the probability of words occurring before and after the input word. Each time the probability of occurrence of the preceding or following word is predicted, the word embedding (which is the hidden layer of this neural network) gets updated.

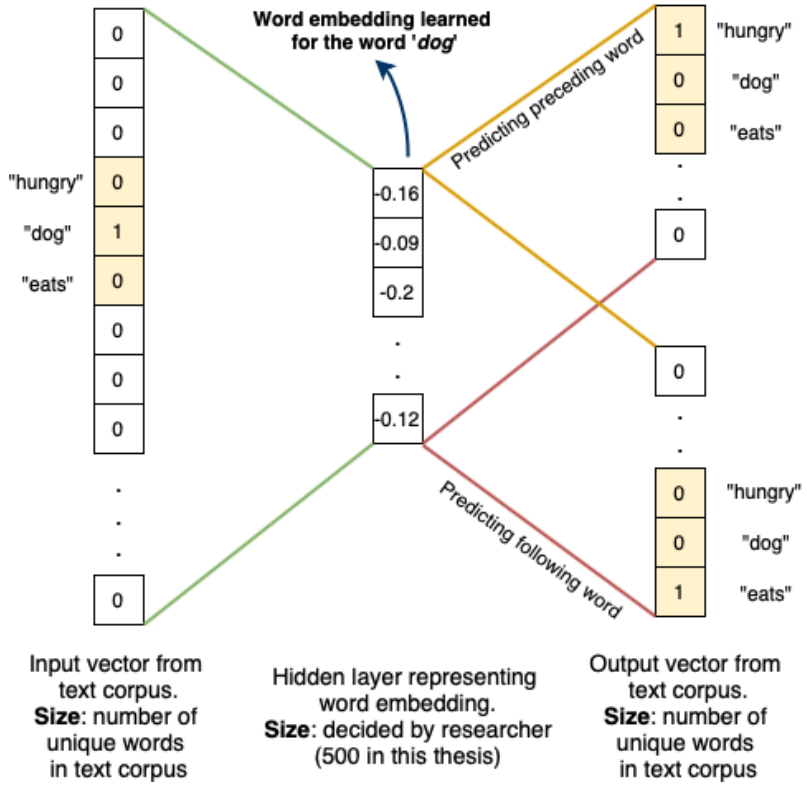


Figure 3.2. *Example of the working of skip-gram architecture used by Word2Vec.*

Intuitively, as shown in Figure 3.2, if the context window in the corpus is "*hungry dog eats*", then the model obtains the word embedding for the target word dog by updating

the embedding with the probability that the words *hungry* and *eats* occur in the same context as the word *dog*. Upon training this model over a large enough corpus, embeddings of concepts such as *cat* and *dog* would end up closer to each other since words such as *eats*, *sleeps*, *pet*, etc. would co-occur for both these concepts. This activity would result in a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector or embedding in this space. The embedding for each concept is usually stored as a row vector in a matrix representing the linguistic vector space (see Figure 3.3). Similarity measures, such as the cosine similarity metric used in this thesis, can then be used to calculate the linguistic similarity between any two concepts per the linguistic model. Here, cosine similarity can be defined as the cosine of the angle between two non-zero vectors, which is calculated as:

$$\cos(\theta) = \frac{\mathbf{A}.\mathbf{B}}{||\mathbf{A}||\,||\mathbf{B}||}$$

where θ is the angle between vectors **A** and **B** (say vector representations of two concepts *dog* and *cat*), between which the similarity is being calculated. In other words, the cosine similarity between vectors A and B is the ratio of their dot product to the product of their magnitudes.

For the purpose of this thesis, the word embeddings for all the concepts were obtained using one-shot learning. One-shot learning is a widely used optimisation technique in machine learning where feature vectors generated upon training complex models such as Word2Vec are saved and made publicly available with the intention of using them in different but related problems. Compared to traditional machine learning approaches, this technique saves upon the vast compute and time resources needed to train complex models so that researchers can apply the available concept representations to their specific use-cases.

This thesis makes use of one-shot learning by acquiring the word embeddings generated by Rei and Briscoe (2014) for all unique words in the British National Corpus and made publicly available for academic purposes. Rei and Briscoe trained their Word2Vec-based neural network on over 100 million words from the British National Corpus. Each distinct word in the corpus was projected into a 500-dimensional vector space. The embedding size of 500 was convenient since these embeddings could be loaded and used in memory for all the unique words in their training data without issues with the memory capacity. Their publicly available embeddings dataset is used to obtain linguistic feature vectors for the object concepts in this thesis, each of which is stored as a row vector in a matrix representing the linguistic vector space. A subset of the resulting feature matrix is shown in Figure 3.3.

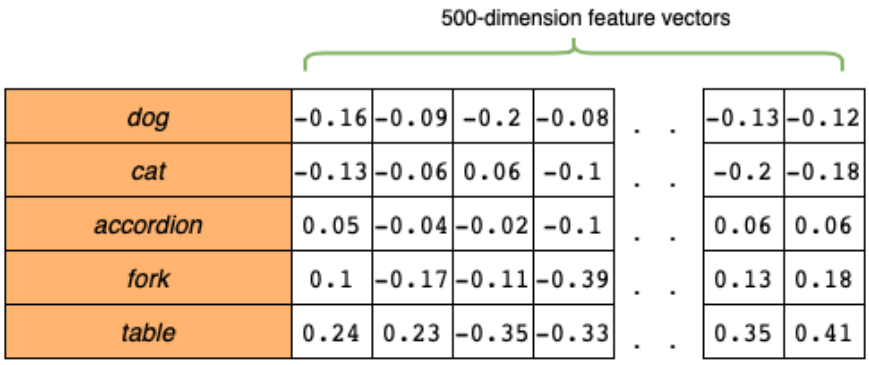


Figure 3.3. *Example of linguistic feature vectors stored row-wise in a matrix.*

Turney and Pantel (2010) assert that since distributional semantic models extract knowledge automatically from a given corpus, they require much less labour and cover multiple times more concepts than approaches relying on hand-coded features. Such models have been shown to perform well on tasks that involve measuring the similarity of meaning between words, phrases, and documents (Lin and Pantel, 2001; Turney, 2006; Nakov and Hearst, 2008; Manning et al., 2008). Due to their considerable success in simulating human behaviour, these models have previously been used for stimuli selection in semantic priming tasks (Bullinaria and Levy, 2012; Padó and

Lapata, 2007). In the next chapter, I present the utility of this model in the production of confusable object pairs and assess its performance using standard evaluation procedures.

3.2.2 Deep convolutional neural networks for visual similarity

Deep convolutional neural networks (Deep-CNNs) are machine learning models of visual object recognition that have revolutionised the field of computer vision in recent years. These deep neural network models are trained on large, labelled image datasets such as ImageNet (Deng et al., 2009) to learn efficient representations of coloured real-world images and have been claimed to reach human-level performance on image classification tasks (Krizhevsky et al., 2012; Szegedy et al., 2016). These models have often been compared and shown to work similar to how the mammalian visual cortex works (see Yamins et al., 2014; Kriegeskorte, 2015; Rajalingham et al., 2018). Specifically, the progressive encoding of an image in deep-CNNs (as shown in Figure 3.4) (Zeiler and Fergus, 2013), i.e., the detection of edges and primitive shapes in the earlier layers and their integration to form more complex visual shapes in the later layers, has been shown to be similar to the sequence of processing stages in the primate visual system (Serre et al., 2007). Similarly, Lee et al. (2007) found that the features learned in deep-CNN architectures resemble those observed in the first two areas (V1 and V2) of the visual cortex. Last but not least, fMRI studies (Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and van Gerven, 2015) have shown similarity between representations of object images in the high-level visual cortex and the later layers of deep-CNNs. Therefore, it is now believed that these deep network models might be able to predict human behaviour in cognitive tasks such as judging object similarity (Jozwik et al., 2017) upon being presented with images of objects.

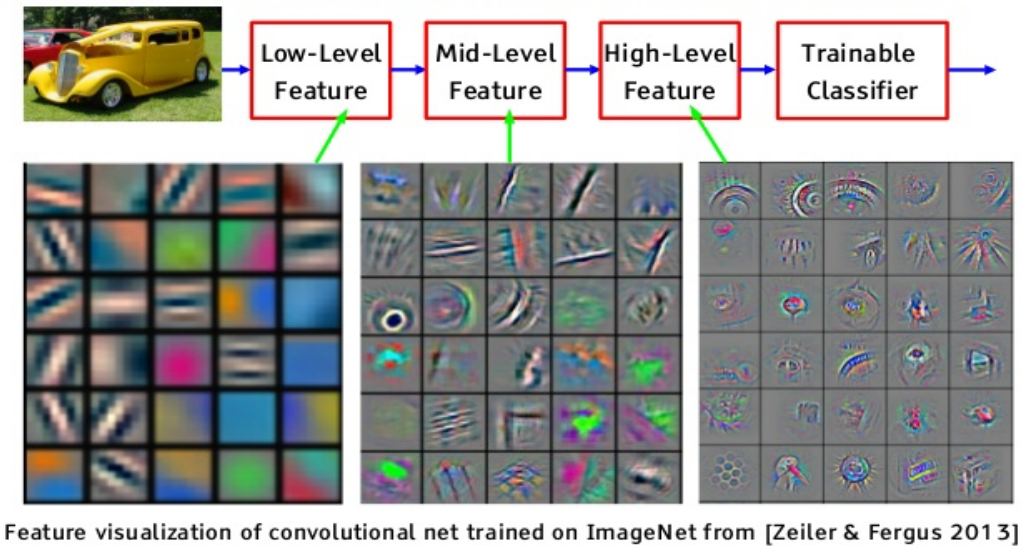


Figure 3.4. *Visualisation of progressive encoding of image features in deep-CNNs (courtesy: Canziani and LeCun, 2020).*

Inspired by these claims, one of the best performing deep-CNN models of image classification on the ImageNet Large Scale Visual Recognition (ILSVR) challenge 2015 (Russakovsky et al., 2015), the Inception-v3 (Szegedy et al., 2016), has been used in this thesis to obtain visual feature vectors for all the object images used in this thesis. The ILSVR challenge recognised the best computer vision models of object detection and image classification each year from 2012 to 2017, thus helping researchers track the progress in the field. The working of the Inception-v3 model is briefly explained next.

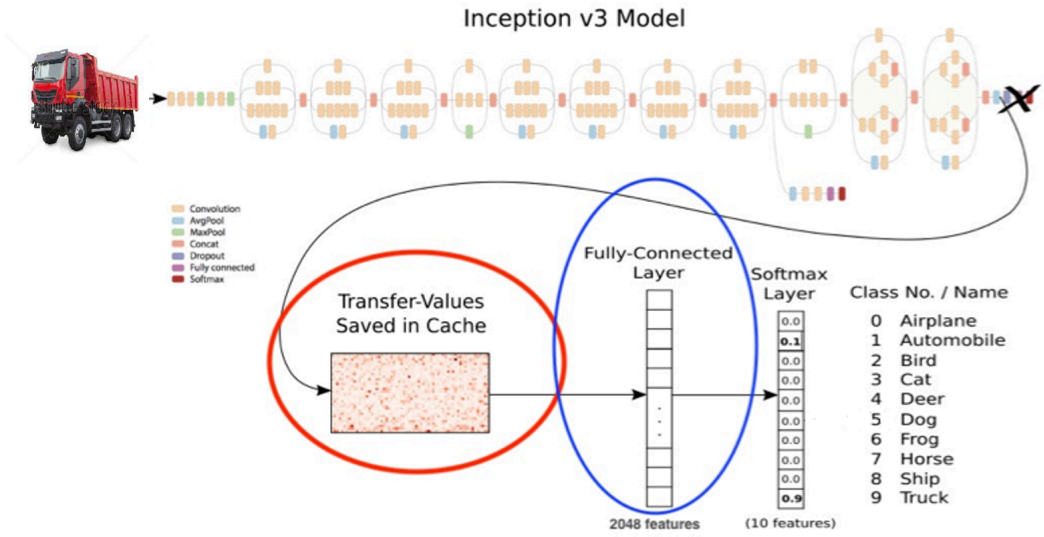


Figure 3.5. *Architecture of the Inception-v3 deep-CNN model (adapted from Szegedy et al., 2016).*

Inception-v3 is a widely used image recognition model claimed to have achieved 78.1% accuracy on the ImageNet dataset. A schematic diagram of the architecture of Inception-v3 deep-CNN model is shown in Figure 3.5. As shown, the model is composed of blocks of convolution, pooling, concatenation, dropout, and fully connected layers. The model works by taking an image as input and convolving "filters" over it in a sliding window fashion. Here, filters are learnable weights as with any neural network (spatial equivalents of hidden layer neuron weights in traditional neural networks as shown in the previous section). By sliding these filters over the image spatially, the model gradually learns image features to produce activation maps. Three such activation maps are shown in Figure 3.4. While activation maps produced in earlier layers contain only simple features such as lines and edges, they transfer these features to the later convolution blocks, where these add up to produce more and more complex feature maps in a hierarchical fashion. For example, in Figure 3.4, the first feature map contains simple lines and edges, whereas the final feature map contains extended parts of objects such as wheels, car grille, etc. The convolution blocks are supported by the other blocks (pooling, concatenation, and dropout) in

thresholding, selecting and overlapping features block after block until the final, fully connected layer, which contains the final feature vector of the input image as shown in the bottom half of Figure 3.5.

The image feature vectors representing object images in this thesis are obtained from this fully connected layer of the Inception-v3 model marked in blue in Figure 3.5. Each of these feature vectors contains 2048 features. Similar to how a vector space of linguistic features was shown to be constructed in the previous section (Figure 3.3), visual feature vector space is constructed using these image vectors. This vector space is then used to calculate the visual similarity between respective image pairs using cosine similarity, which is a representation of the visual feature overlap between such pairs of images. The thus obtained pairwise visual similarities can then be used in compiling the intended database of highly confusable image pairs. The implementation and evaluation of this model is discussed at length in the next chapter.

3.2.3 Lexical Semantic Networks for taxonomic similarity

The taxonomic source of computational information used to measure similarity between object pairs in this thesis is the WordNet lexical semantic network. Lexical semantic networks are networks of words interlinked with a variety of relations. The WordNet database is a lexical database that offers a taxonomy of a vast range of word concepts (155,327 as of July 2021) organised in a hierarchical tree-like structure according to their meaning and semantic relationship with other concepts. Specifically, concepts are organised such that every child node is a hyponym (IS-A relationship) of the root node. An example sub-tree within WordNet is shown in Figure 3.6. Each node of this tree is called a “synset”, which is a WordNet identifier that makes it easy not only to find the relationship between concepts (such as similarity, meronymy, etc.), but also to disambiguate between different senses of the same word. For instance, the

synset “crane.n.04” is defined in WordNet as “lifts and moves heavy objects; lifting tackle is suspended from a pivoted boom that rotates around a vertical axis”, whereas the synset “crane.n.05” is defined as “large, long-necked wading bird of marshes and plains in many parts of the world”. WordNet is available for the Python programming language through the NLTK library (https://wordnet.princeton.edu/).

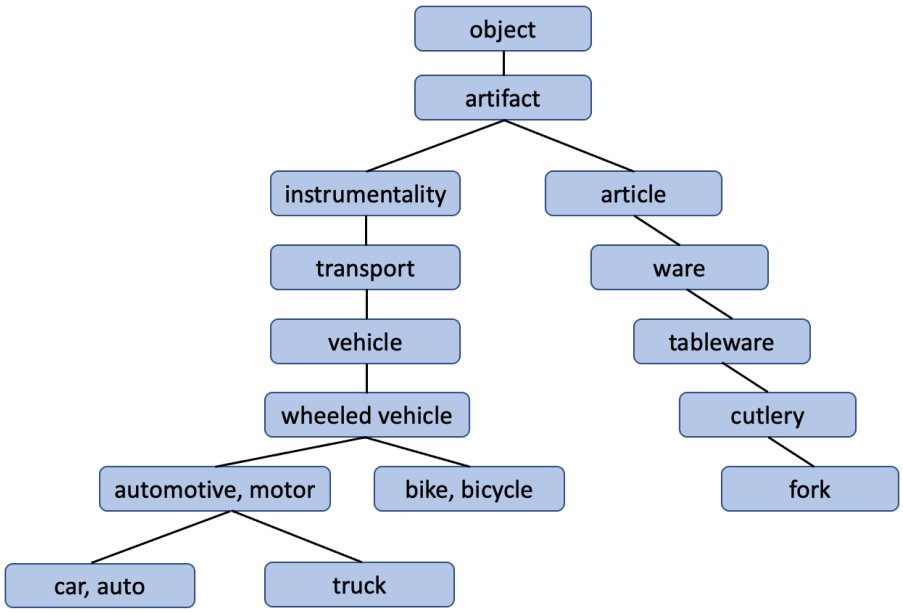


Figure 3.6. *Example WordNet sub-tree*

Having a tree-like structure with lexical relationships between concepts also makes it possible to compute the semantic similarity between them. Simple measures such as path length between two concepts (Rada et al., 1989) are widely used to determine the semantic similarity between any two concepts (the shorter the path length, the more similar the concepts). However, such edge counting methods are criticised to be too simplistic since they rely on the assumption that “links in the taxonomy represent uniform distances” (Resnik, 1995 in Budanitsky and Hirst, 2006, p.16), which has been shown to not be true in WordNet, since certain sub-trees such as biological sub-taxonomies are much denser (i.e., have more child nodes at each level of the taxonomy) than others (see Sussna, 1997). To address this shortcoming, normalised and scaled measures of semantic distance (Sussna, 1997; Wu and Palmer, 1994;

Leacock and Chodorow, 1998) are now widely used (see Budanitsky and Hirst, 2006 for a review of these measures). Among these, the most used technique for determining concept pair similarity is Wu and Palmer's similarity metric (Wu and Palmer, 1994; hereafter referred to as WUP similarity metric).

The WUP similarity metric is defined as:

$$\text{sim}_{\text{WUP}}(c_1,c_2) = \frac{2*depth(lso(c1,c2))}{len(c1,lso(c1,c2))+len(c2,lso(c1,c2))+2*depth(lso(c1,c2))}$$

where:

1. c1, c2 refer to the two concepts (or words) between which similarity is being calculated,
2. $lso(\text{c1},\ \text{c2})$ refers to the lowest superordinate or the most specific common subsumer (e.g., 'mammal' is the *lso* for '*cat*' and '*dog*') of c1 and c2, and
3. $depth(lso(\text{c1},\ \text{c2}))$ is the depth of the *lso* from the "global" root of the WordNet tree.

The WUP similarity metric corrects the shortcoming of simpler edge counting approaches by normalising the distances between concepts c1 and c2 with respect to their most specific subsumer (*lso*). In this thesis, the WUP similarity metric has been used to calculate the taxonomic similarity between any two object concepts, thereby helping create the intended database of confusable object pairs.

3.2.4 Similarity measures for vector space models

Similarity measure may be defined as a real-valued function that quantifies the alikeness between two objects. In the context of vector space models, similarity is

usually quantified as the inverse of distance metrics, i.e., if the distance between the vector features (dimensions) is small, the two vectors are very similar and vice versa.

Properties of a similarity measure:

1. Symmetry: sim(**A**, **B**) = sim(**B**, **A**) for **A** and **B**, where sim(**A**, **B**) is the similarity between vectors **A** and **B**.
2. Maximum similarity: sim(**A**, **B**) = 1 (or maximum similarity) only if **A** = **B.**
3. Non-negativity: In the context of distance (i.e., as explained above, similarity is the inverse of distance), dis(**A, B**) $\geq$ 0 for vectors **A** and **B**. While the property of non-negativity also applies to similarity, i.e., sim(**A, B**) $\geq$ 0, in some metrics the similarity measure is also dictated by the direction of the vectors, sim(**A**, **B**) = -1 if the vectors are opposite to each other. Therefore, only the magnitude of similarity |sim(**A**, **B**)| $\geq$ 0 for vectors **A** and **B** in such measures.
4. Triangle inequality: If vectors **A** and **B** form two sides of a triangle, where the third side is the sum of the two vectors, then the triangle inequality states that for vectors **A** and **B**, ||**A** + **B**|| $\leq$ ||**A**|| + ||**B**||, i.e., the norm of the sum of two vectors is at most as large as the sum of the norms of the two vectors. A natural progression to the triangle inequality is the reverse triangle inequality, which states that any side of a triangle is greater than or equal to the difference between the other two sides, i.e., ||**A** - **B**|| $\geq$ | ||**A**|| - ||**B**|| |.

Examples of common similarity measures: Among the proximity measures of data objects, the following five measures are the most used in the realm of machine learning: Manhattan distance, Euclidean distance, Minkowski distance, and cosine similarity. These are discussed in detail next:

1. Manhattan distance: The Manhattan distance is the sum of the absolute difference between the features (or dimensions) of two vectors. In other words, it is the L1 norm between two vectors. It is represented as dis(**A**, **B**) = $\sum_{i=1}^{n}(|\boldsymbol{A}_i - \boldsymbol{B}_i|)$, where *n* is the number of dimensions of the vectors **A** and **B**. This metric is more useful to calculate distance between vectors that describe objects in a uniform grid, like a chessboard or city blocks. For this reason, this metric is usually used to calculate the distance (and hence the similarity) between vectors in an integer feature space.
2. Euclidean distance: The Euclidean distance is the L2 norm or the sum of squared distance between the features of two vectors. It is represented as dis(**A**, **B**) = $\sum_{i=1}^{n}\sqrt{(\boldsymbol{A}_i - \boldsymbol{B}_i)^2}$, where *n* is the number of dimensions of vectors **A** and **B**. In addition to vectors with integer values, this distance metric can also be used to calculate the distance between vectors with floating point or integer values. If the corresponding features in the two vectors have values with differing scales, they must be normalised or standardised prior to using the Euclidean distance measure to ensure that features with large values do not dominate the distance measure.
3. Minkowski distance: Generalising the formulae for the previous two distance measures, we get the Minkowski distance dis(**A**, **B**) = $\sum_{i=1}^{n}\sqrt[p]{(\boldsymbol{A}_i - \boldsymbol{B}_i)^p}$, where *p* is called the "order parameter". When set to 1, it gives the Manhattan distance, and when set to 2, it gives the Euclidean distance. Intermediate values provide a controlled balance between the two measures. When using this metric in machine learning problems, the hyperparameter *p* can be tuned to maximise the underlying objective of the problem.
4. Cosine similarity: For vectors **A** and **B**, cosine similarity is the cosine of the angle between two non-zero vectors, calculated as:

$$\cos(\theta) = \frac{\mathbf{A}.\mathbf{B}}{||\mathbf{A}||\,||\mathbf{B}||}$$

where θ is the angle between vectors **A** and **B**. In other words, the cosine similarity between vectors A and B is the ratio of their dot product to the product of their magnitudes. Cosine similarity $\in$ [-1,1]: for example, for two proportional vectors, cosine similarity is 1, for two orthogonal vectors, cosine similarity is 0, and for two opposite vectors, the cosine similarity is -1. Cosine similarity is used in numerous machine learning applications including but not limited to document classification and computer vision, where the interest lies more in the similarity irrespective of the sizes of the input vectors.

Rationale for the choice of the cosine similarity measure: As stated before, the choice of the similarity metric used in this thesis is the cosine similarity. This choice was made in line with the similarity metric used in past literature for multidimensional vector spaces, where there are three key reasons for the choice of this metric. First, compared to the other measures described above, the cosine similarity metric does not use the magnitude of the vectors, but rather relies on the direction of the vectors to measure similarity. This is more appropriate when it is likely that the documents or images may be of different sizes, but we are interested in their similarity irrespective of their sizes. Second, cosine similarity is considered to be the more accurate metric of similarity between two multidimensional vectors because while cosine similarity is scale invariant, it is not affine invariant. This means that if we add or subtract a constant from a feature (or dimension) in the multidimensional vector, its cosine similarity will change, while the other distance measures will remain the same. This means that changing data across any dimension will impact the relationship between the two vectors, which is a useful relationship to capture when measuring similarities between two image or word data objects. Finally, in sparse matrices, such as the one

we obtain from the deep learning models in this thesis, cosine similarity is faster to calculate.

3.3 Production of confusable object pairs: The Hungarian algorithm

The central aim of this thesis is to use the object pair similarities obtained from computational models to generate datasets of the most optimally matched target-distractor pairs from each model as per the following criterion: each obtained target-distractor pair should be unique in the generated dataset such that neither the target nor the distractor should be paired with another concept in the dataset. As discussed in Chapter 1 (section 1.6), this criterion is necessary to produce only one distractor image per target image and vice versa given the setup of memory assessment tasks discussed in the previous chapters. For instance, assuming that a model outputs the concept *tiger* as the most similar image for the concept *lion*, then neither the concept *tiger* nor *lion* could be paired with any other concept. This rigidity in selecting unique pairs of concepts, however, presents a challenge: due to the non-availability of the most similar distractor concept, a target concept would be paired with the second most or third most similar concept. For instance, once the concept *tiger* has been paired with the concept *lion*, despite another concept *jaguar* having the highest pairwise similarity with the concept *tiger* as well, it might have to compromise with the second most similar concept to it (for instance, concept *panther* in Figure 3.7).

	lion	*tiger*	*Jaguar*	*panther*
lion	1	**0.79**	0.69	0.65
tiger	**0.79**	1	0.81	0.72
jaguar	0.69	0.81	1	**0.89**
panther	0.65	0.72	**0.89**	1

Figure 3.7. *Values in cells show the example similarities between two concepts. Hungarian algorithm selects the values resulting in net maximum similarity index from the matrix.*

To address the problem whether the concept *tiger* should be paired with the concept *lion* or *jaguar*, where both the latter concepts individually have the highest similarity index with the concept *tiger*, a choice has been made in this thesis to select pairs that maximise the sum of all object pair similarities in each obtained database. Therefore, if pairing *tiger* with *lion* resulted in a higher net sum of object pair similarity values in a database, then the algorithm would make such a choice (see Figure 3.7). To solve this problem, the Hungarian algorithm (Kuhn and Yaw, 1955), also called the Munkres algorithm (Munkres, 1957) is used.

	J1	*J2*	*J3*	*J4*
W1	82	83	**69**	92
W2	77	**37**	49	92
W3	**11**	69	5	86
W4	8	9	98	**23**

Figure 3.8. *Hungarian algorithm selects the values corresponding to net minimum cost of assigning "**j**" jobs to "**w**" workers (courtesy: http://www.hungarianalgorithm.com/).*

The Hungarian algorithm is a combinatorial optimisation algorithm that solves the assignment problem in polynomial time, where the assignment problem can be stated as follows: given *j* jobs that need to be executed by *w* workers (one job per worker) with a cost matrix showing the cost of assigning a certain worker to a certain job, the goal is to minimise the total cost of the assignment. Figure 3.8 shows such optimal values selected in order to minimise the total cost. Translating this to the problem at hand, given one distractor object to be selected for one target object, with a similarity matrix showing the similarities between all pairs of objects, the goal is to maximise the total similarity of all the unique pairs obtained. As can be observed, the key difference between the problem of choosing most similar concept pairs from a pairwise similarity matrix and the assignment problem is that while the assignment problem is an attempt at *minimising* the total cost of the assignment, the problem of maximum pairwise similarity is to *maximise* the total similarity of all obtained pairs.

To reflect this difference, only one change has been made to the Python implementation[i] of the Hungarian algorithm in the present thesis to obtain the required pairs – all the pairwise similarity indices in the similarity matrix are negated before inputting them in the algorithm. This way, a minimum cost calculated by the Hungarian algorithm maximises the negative of the sum of object pair similarities in the matrix. As a result, the most appropriate pairs per model can be obtained and a database of confusable object pairs with varying degrees of similarity can be constructed. A detailed explanation of the implementation of Hungarian algorithm is available online[ii].

3.4 Evaluation techniques

To quantitatively evaluate the computational models on their ability to predict similarity between concepts, studies have historically relied on assessing how well such models mimic human understanding of concept pair similarity. For this purpose, two key methods of evaluation have been used in existing literature. The first method measures the proportion of all object pairs produced by the model in which the constituent pairs of objects belong to the same semantic category. The second method compares the model-produced concept pair similarity ratings with their human-rated counterparts, for which methodically acquired datasets of human-rated concept pair similarities are used. Both these techniques are discussed in detail below.

3.4.1 Assessment on semantic category membership

A number of studies (e.g., Riordan and Jones, 2010; Fountain and Lapata, 2010; Baroni et al., 2014; Silberer, 2015) have previously used semantic category memberships of objects rated similar by computational models to measure the

[i] The scipy library provides an implementation of the Hungarian algorithm with the function scipy.optimize.linear_sum_assignment, which takes as an input a cost matrix and returns unique pairs with one-to-one mapping such that pairs are maximally matched.

[ii] http://www.hungarianalgorithm.com/ gives a step-by-step explanation of the implementation of the algorithm.

effectiveness of such computational models at mimicking human behaviour. In the context of this thesis, this metric assumes that given an object dataset and the problem to pair the most similar object images, humans would try to maximise the number of pairs in which objects belonged to the same semantic category. As a reminder, one of the traits of highly confusable object pairs listed in Chapter 1 is that the constituent objects from such pairs belong to the same semantic category. As such, category-based evaluation of models in this thesis is done by measuring the proportion of all object pairs produced by each model in which the constituent objects belong to the same semantic category. The higher this metric, the more effective the model. As an example, both concepts in the pair <*cow, horse*> belong to the same semantic category (MAMMALS), whereas concepts in the pair <*snake, rope*> do not belong to the same semantic category (REPTILES, TOOLS).

For the evaluation of the models implemented in this thesis on the semantic category membership metric, the object pairs from each model were produced using the Hungarian algorithm and the semantic categories were adapted from Hebart et al. (2019). Choosing semantic categories has been a widely debated topic in the field of object classification majorly due to a lack of agreement on the levels of abstraction needed to suitably segment real-world objects into these categories. For instance, is an exemplar of a *dog* suitably defined as a type of concept *dog*, concept *mammal* or concept *animal*? Various attempts have been made to formalise the principles of object categorisation over the years (see Rosch and Lloyd, 1978; Neisser, 1987; Tanaka and Taylor, 1991; Adlington et al., 2009 and Bauer et al., 2017 for the attempts at this formalisation). These attempts have resulted in the separation of concepts into three broad levels of categorisation, namely superordinate, basic and subordinate-level categories. Superordinate level categories are placed at the top of a typical taxonomy, display a high degree of generality, and provide very abstract information about the category (e.g., concept ANIMAL is a superordinate level

category). Basic level categories are defined by specific attributes that are common to all or most members of the category (e.g., concepts MAMMAL, BIRD and FISH are basic level categories). Finally, subordinate level categories are defined by features that, like basic level, are also common to members of those categories, but may also be common with other subordinate categories (e.g., concepts *dog* and *cat* are subordinate level categories).

Different researchers, however, place different concepts in each of these categories. As a result, the number of categories at the basic level, for instance, may vary from one research piece to another. For example, among some of the widely known research pieces published in this area in just the last decade and a half, the Almuhareb-Poesio benchmark used 21 semantic categories (402 concepts: Almuhareb, 2006), ESSLLI 2008 Distributional Semantic Workshop used 6 semantic categories (44 concepts: Baroni et al., 2008) and Battig test (83 concepts: Baroni et al., 2010) used 10 categories. However, given that the concept dataset used in this thesis (shared in the next chapter) was much larger than these (1,402 concepts), experimentally determined semantic categories from a comparable concept dataset (Hebart et al., 2019: 1,854 concepts) were chosen and minimally adapted to align with the object concepts compiled in this thesis. These are listed in Table 3.1 below. Each concept in the dataset used in the present thesis was labelled with its semantic category in-house and independently verified by two researchers. Upon obtaining these category labels, the object pairs produced by each model for this dataset were subject to the semantic category membership method of evaluation.

Table 3.1. *29 semantic categories of objects used in this thesis.*

ACCESSORY	APPLIANCE	BIRD	CLOTHING	CRUSTACEAN
DEVICE	FISH	FLOWER	FOOD	FRUIT
FURNITURE	HOUSEHOLD	INSECT	INSTRUMENT	MAMMAL
MISCELLANEOUS	MOLLUSC	MUSIC	PLANT	REPTILE
RODENT	SPORT	STATIONERY	STRUCTURE	TOOL
TRANSPORT	UTENSIL	VEGETABLE	WEAPON	

3.4.2 Comparison with human judgements of concept pair similarity

While the semantic category membership metric has been widely used to evaluate computational models, it has a critical limitation: this metric does not give an indication of the within-category similarity of concept pairs. For instance, a *tiger* and a *goat*, both belong to the same superordinate category MAMMALS and according to this metric, if a <*tiger, goat*> pair is generated by a model, then it would contribute to the effectiveness of a model per this metric even though the two concepts are visually dissimilar and semantically more distant as compared to, for instance, <*tiger, lion*>.

To address this shortcoming, another method often used by studies to assess the performance of computational models is to evaluate the performance of such models on existing human-rated datasets of concept pair similarity (see Budanitsky and Hirst, 2006). In such a setting, computationally produced similarities between chosen concept pairs are compared with the human similarity ratings for those concept pairs. This comparison is facilitated in previous research using Spearman's rank correlation (see Hill et al., 2015). The thus obtained Spearman's rank correlation indicates the extent to which a model's similarity ratings are aligned with the human-produced similarity ratings for the same set of object pairs. Since the aim of using these models in this thesis is to produce parametrically varying object pairs, this metric is a suitable

indicator of a model's effectiveness at producing such object pairs per the human-rated dataset it is evaluated against.

Several experiments (Rubenstein and Goodenough, 1965; Miller and Charles, 1991; Finkelstein et al., 2001; Bruni et al., 2014; Silberer and Lapata, 2014; Hill et al., 2015) have been conducted to obtain pairwise ratings of concepts from human participants with the intention of evaluating computational models on these pairs. Here, I elaborate on two popular datasets chosen for evaluation of the models in the next chapter: 1) Semantic and Visual Similarity Judgements for Concept Pairs (Silberer and Lapata, 2014) and 2) SimLex-999 (Hill et al., 2015). Among the datasets cited earlier in this section, these datasets were chosen because during their compilation, the participants were explicitly asked to rate the concept pairs on similarity, while the earlier datasets were mostly rated on association. Similarity (and not association) between objects in a pair has been listed as a key requirement in the characteristics of ideal object pairs in section 1.6 of this thesis. Additionally, both these datasets contain shared concept pairs with the database compiled in this thesis. However, since many of the concept pairs included in these existing datasets are abstract and non-picturable (e.g., <*music, sound*>), they will be absent from this analysis. Despite this, past studies have shown that even a subset of concept pairs provides a good approximation of model effectiveness (Lazaridou and Baroni, 2015; Faruqui et al., 2015; Derby et al., 2018). A brief description of the two datasets is as follows.

Semantic and Visual Similarity Judgements for Concept Pairs (SemSim and VisSim): Silberer and Lapata (2014) presented a dataset of 7,576 concept pairs rated for similarity by human participants in a study hosted on the Amazon mechanical Turk platform. This database is important for the evaluation of models in this chapter for two reasons: a) It covers exclusively concrete noun concept pairs obtained from McRae et al. (2005), leading to 7,576 human-rated concept pairs, and b) it contains separate

ratings for both visual and semantic similarity of the said object pairs, thus evaluating two key dimensions of similarity considered in this research. A total of 3,838 concept pairs were found overlapping with the database used in this thesis.

SimLex-999: This state-of-the-art word similarity dataset by Hill et al. (2015) provides similarities between 999 concepts pairs, rated by 500 human participants via the Amazon mechanical Turk platform. The authors have shown that this dataset is different from previous gold standards of human ratings of similarity between concepts (Finkelstein et al., 2001; Bruni et al., 2014) in that while compiling this database, the authors clarified the difference between similarity and association to the participants by relating it to synonymity of objects (for instance, $Sim_{(bread,\ toast)} > Sim_{(butter,\ bread)}$) and explicitly asked them to rate the pairs on similarity rather than association. The authors have shown through analysis that this disambiguation between object similarity and association ensured that the participant ratings reflected similarity and not association unlike many of the previous such datasets. Of note, SimLex-999 consists of similarity ratings for non-picturable nouns, verbs, and adjectives, and only 54 human-rated concept pairs could be found from the study that overlapped with the database used in this thesis.

3.5 Conclusion

In this chapter, I provided a background of three major strands (visual, linguistic, and taxonomic) of computational models, introduced the algorithm that can be used to produce datasets of unique confusable object pairs from these, and summarized the methods of evaluation of such models. The visual and linguistic models were introduced as vector space models that project object representations in n-dimensional vector spaces, which can be used to obtain relationships between objects upon the application of suitable mathematical techniques. Contrastingly, a tree-based model

represented the taxonomic model, in which the relationships between object pairs can be obtained using optimized path length techniques. Following this, the Hungarian algorithm was introduced as a suitable technique for creating datasets of unique object pairs from model-generated object pair similarity values. Lastly, model evaluation techniques used in previous literature were discussed and the semantic categories and the human-rated datasets of object pair similarities needed for model evaluation were shortlisted from existing literature.

Chapter 4

Production of confusable object pairs using computational models

4.1 Introduction

In the previous chapter, I provided a background of the computational models of vision and semantics and how they can be used to capture representative features of each object. In the present chapter, I show how these models and their weighted combinations were used to produce datasets of confusable object pairs parametrically varying in similarity in this thesis. I also evaluate the produced datasets using standard techniques discussed in the previous chapter. With the help of this evaluation, I facilitate a comparison among the implemented models. Such a comparison would aid in the selection of confusable object pairs produced by the best performing models for downstream validation and use in memory assessment tasks presented in later chapters. To perform such analyses, as well as to produce the required confusable object pairs, however, there was first a need to assemble a large dataset of objects and their representative images.

In light of the above, I first present how I assembled a dataset of object names and their representative images in section 4.2. Following this, in section 4.3, I present the pipelines designed to produce confusable object pairs using the computational models central to this thesis. In section 4.4, I present the statistical analysis conducted to evaluate and compare the models as well as the datasets of confusable object pairs they produced. In section 4.5, I present the findings from this analysis. In section 4.6, I discuss the obtained findings and conclude the chapter in section 4.7.

4.2 Acquisition of the object image database

As mentioned, the production of confusable object pairs is dependent on the availability of a large dataset of images of distinct object concepts. To obtain such datasets of object images, studies in the past have first compiled a list of unique object concepts and then sourced the images representative of such concepts. The problem of assembling the lists of object concepts and their corresponding images is not new and has been addressed systematically in existing research as presented next.

4.2.1 Past methods for acquiring object images

Most existing image databases use Battig and Montague concept lists for the acquisition of their respective images. Battig and Montague (1969) were among the first to systematically address the need for acquiring large databases of concepts for human participant-based studies. To compile such a database, they first set down the semantic categories (e.g., four-footed animals, carpenter's tools, musical instruments, etc.) across which a large number of concepts could be listed. They then organised a data collection activity in which 400+ undergraduate students were asked to list as many concepts as possible for each semantic category in 30 seconds. A total of more than 5,000 concepts (both picturable and non-picturable) were listed in this activity. Snodgrass and Vanderwart (1980) later used these concepts from Battig and Montague (1969) to compile the first popular database of concept images for assessment of memory and cognition. They handpicked 260 picturable concepts from 15 of 56 semantic categories and presented black-and-white line-drawn pictures of these concepts (see Figure 4.1) for use in memory assessment tasks. Brodeur et al. (2010) recognise that these pictures rapidly spread across the scientific community and were widely used in cognitive assessment tasks. Over the years, several studies supplemented this image database with more pictures (e.g., Cycowicz et al. 1997; Bonin et al., 2003; Nishimoto et al., 2005; Álvarez and Cuetos, 2007).

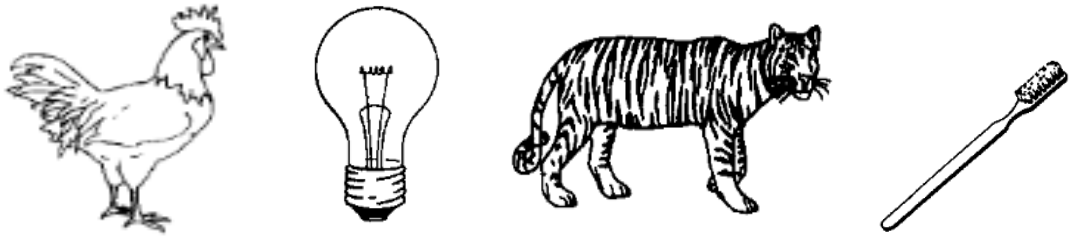


Figure 4.1. *Line drawn pictures from Snodgrass and Vanderwart image dataset (courtesy: Snodgrass and Vanderwart, 1980).*

However, photo stimuli and line drawings are characterised by different features that serve different purposes in object processing and identification. While line drawings are prototypical semantic representations of objects, historically used for evoking a concept (Ostergaard and Davidoff, 1985), coloured photo stimuli come with colour, texture, and 3D cues (such as shade) – variables that have been shown to influence recognition, disambiguation, and naming of objects in memory assessment tasks (Davidoff and Ostergaard, 1988; Brodie et al., 1991; Rossion and Pourtois, 2004). Several studies have shown how change in elements such as texture, form and colour of objects can induce confusability in disambiguation of two different exemplars of the same concept (Yassa et al., 2011b; Stark et al., 2019). Owing to this need, new datasets of coloured isolated images of everyday objects were constructed.

Among the earliest such systematically acquired datasets were the ones made available by Viggiano et al. (2004: 174 photos) and Adlington et al. (2009: the Hatfield Image Test or HIT, 147 photos). In 2010, an even bigger set of 480 normative photos of everyday objects was made publicly available by Brodeur et al. (2010). However, the number of images in these datasets remained small, owing to two key constraints: 1) all these datasets still relied on the concept list first made available by Battig and Montague (1969) and 2) these datasets necessitated name agreement from a large number of people on all object images in their database, i.e., any image that couldn't

be consistently named by a large number of participants in their study was not included in the final dataset. While the former restricted the concept knowledge base of the datasets to Battig and Montague concepts only, the latter ensured that such datasets contained only the most frequently occurring objects that are highly familiar to and precisely nameable by a large population. Scholars, however, agree that while this strictness is beneficial if the images are being used in object naming experiments, for experiments that are using visual stimuli to test mnemonic discrimination, such rules can be relaxed since it is likely that human subjects are using linguistic feature tags to mentally identify the specific exemplars of the objects presented to them (McRae et al., 2005; Hunsaker and Kesner, 2013; Stark et al., 2019). Furthermore, the effects of psycholinguistic variables such as concept familiarity and image complexity corresponding to each stimulus can be adjusted for during the modelling of participant performance on such tasks by including these factors as covariates (as in Montefinese et al., 2015, 2018; Lancaster et al., 2020).

In sum, considering sources beyond the Battig and Montague concept dataset and removing the restriction on the nameability of every image in a database could yield larger image databases. Building upon this discussion, I next present important criteria for the selection of images usable in memory assessment tasks, followed by the creation of a database of 1,402 object images used in this thesis.

4.2.2 Criteria for image selection

Inspired by Snodgrass and Vanderwart (1980), the criteria for image selection for use in memory assessment tasks are listed below:

1. **Non-composite object image:** The image should not be composite, i.e., it should not be a combination of multiple distinct objects and should be an

isolated object (preferably single) in full on a white background. Some examples of difference between composite and isolated object images are shown in Figure 4.2.

2. **Image-object agreement:** The image should represent the target object.
3. **Image relevance:** The image should be a widely agreed upon exemplar of the object.
4. **Familiarity:** The image should be familiar to the general public and should not be a non-existent, made-up object.

(A) **(B)** **(C)**

Figure 4.2. *Composite (A and B) and isolated (C) pictures for the concept 'pineapple'. The images on the left and in the middle are from ImageNet, while the image on the right is from Shutterstock.*

4.2.3 Acquisition of images

The image acquisition pipeline used to assemble the image dataset used in the present thesis is shown in Figure 4.3.

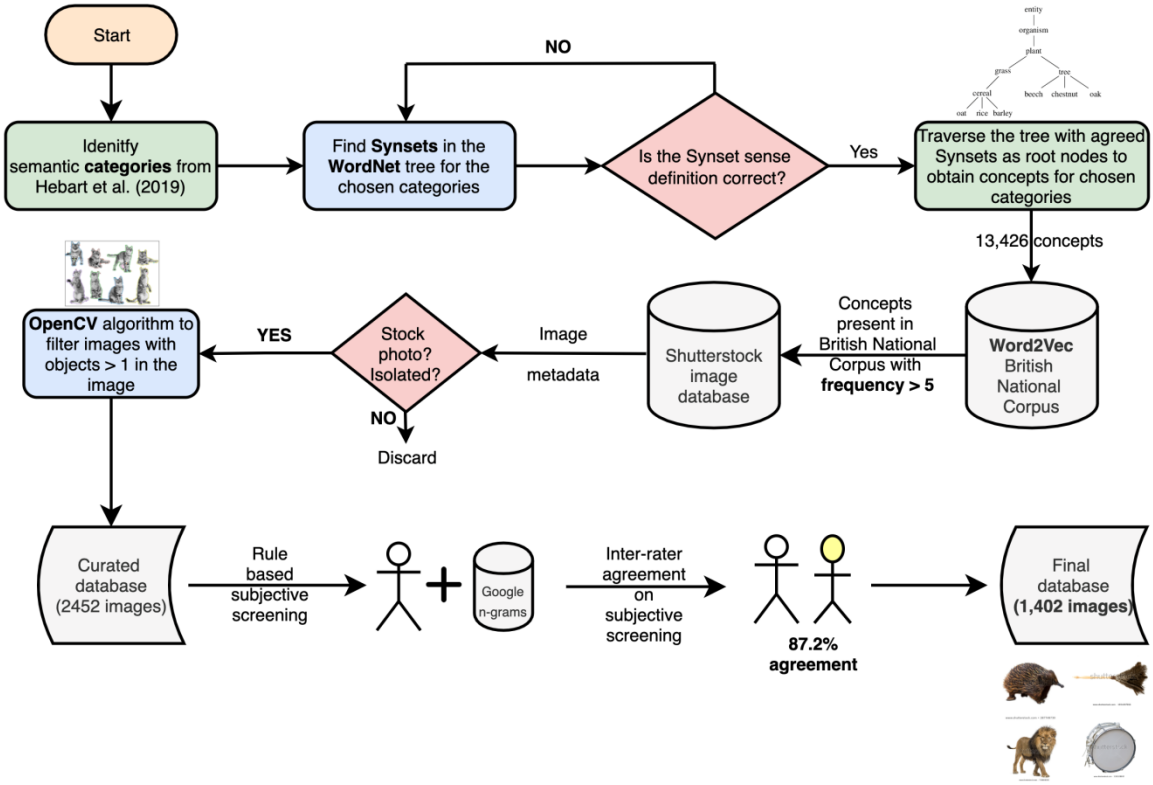


Figure 4.3. *Pipeline for image database acquisition.*

The first step in creating the required database of images was to compile a list of distinct picturable concepts. Instead of the Battig and Montague database, the WordNet database of words in the English language (Fellbaum, 1998) discussed in the previous chapter was used to obtain these concepts. As discussed in the previous chapter, as a lexical database, WordNet offers a taxonomy of a vast range of concepts (155,327 as of July 2021) organised in a hierarchical tree-like structure according to their meaning and semantic relationship with other concepts. Each node of this tree is called a “synset”, which is a WordNet identifier that helps disambiguate between different senses of the same word. To obtain picturable concepts in WordNet, the list of semantic categories adapted from Hebart et al. (2019) and listed in the previous chapter was used. First, the synsets of all the semantic categories (top-level nodes such as ‘mammals.n.01’) were identified in the WordNet tree. These have been shared in the appendix. Upon obtaining such synsets, a tree traversal operation was performed using Python programming language to obtain all concepts in the WordNet

tree under each semantic category used. The operation yielded a total of 13,426 concept names.

As stated in section 4.2.2, an important criterion to account for in concept selection is familiarity. While the benchmark for concept familiarity was not as stringent as in earlier image databases, as a proxy for familiarity, the following rule was imposed to filter concepts that could potentially be too unfamiliar to a large number of people in the English-speaking world: any concept occurring five times or less in the British National Corpus (Andersen et al., 2008) was removed from the concept list. The British National Corpus (BNC) is a 100-million-word text corpus of samples of written and spoken English from a wide variety of sources. The assumption behind this filtering rule was that the corpus was a representation of the knowledge of the English language among general public, and any concept occurring five times or less in this large corpus could be considered unfamiliar to the general public (also see Rei and Briscoe, 2014). Once the concept list of the remaining concepts was compiled, the next step was to acquire images corresponding to such concepts.

To comply with the criteria listed in section 4.1.2 for acquiring images suitable for use in memory assessment tasks, there was a need for a source of images that could be queried to obtain single and isolated pictures of objects on a white background. Academic sources such as ImageNet (Russakovsky et al., 2015) are not suitable for this purpose since images in such datasets are composite and noisy as exemplified in Figure 4.2. Therefore, Shutterstock, a website licensing stock photographs of everyday objects, was used as the source of image acquisition in this pipeline. Isolated pictures with white backgrounds could be downloaded in batches via Shutterstock's cloud-based application programming interface (https://api.shutterstock.com/) and the metadata associated with each picture (e.g., image caption), provided by approved

human content creators and verified by website moderators, was used to validate the remaining criteria listed in section 4.2.2.

While using the Shutterstock API to download images, three key parameters “query text”, “people number” and “image type” could be regulated to obtain the required images. Specifically, including the keyword “isolated” in the query text (e.g., ‘*cat* isolated’) would help download isolated images of the concept; using the value “0” for the attribute “people number” ensured that no humans were present in the images, which could otherwise be possible (as shown in Figure 4.2), and selecting “image type” as “photo” ensured vectors or cartoon drawings as shown in Figure 4.4A were not downloaded.

Figure 4.4. *Unsuitable image examples (A) Vector images for concept 'cat'; (B) cropped image for concept 'cat'; (C) multiple objects in an image.*

Furthermore, the image caption associated with each downloaded image was used to satisfy criteria 2 (image object agreement) and 3 (image relevance) listed in section 4.2.2. Specifically, if the image caption contained the concept name, picture-name agreement was assumed. On the other hand, to ensure that the image was a generally acceptable representation of the concept, the website’s “popularity” ranking feature that returns the most downloaded images for that search query was employed. This was a suitable proxy for the image relevance rule as later confirmed upon manual inspection of the image database by two independent reviewers. To test that the downloaded images conformed to the requirements, the image processing libraries

scikit-image and OpenCV were employed to detect and remove images that had cropped or multiple objects and non-white backgrounds. Along with the code for the rest of the pipeline, the code for this step is also shared in the git repository https://github.com/deveshbatra/confusable-pair-selection in the “ImageApprover” class.

After code-based screening, images were grouped into their semantic categories and two sets of subjective screening were performed: 1) images with different breeds of animals (such as *bulldog*, *chihuahua* and *dachshund* for *dogs*) were removed, keeping only the most popular exemplar (here, *dog*) from such categories (see Hebart et al., 2019 for inspiration); 2) images with no clear distinguishing features with another image in their semantic category were removed based on their frequency of occurrence in the British National Corpus as a proxy for their popularity. The frequency of their occurrence was obtained from Google N-gram model (Google, 2012; 'Google Ngram Viewer'; https://books.google.com/ngrams/datasets). For instance, images of “*alligator*” and “*crocodile*” were indistinguishable, and since “*alligator*” was 3 times less popular as per the Google N-gram model (as shown in Figure 4.5), it was removed from the dataset. These subjective screenings were necessary to obtain a dataset that consisted of images that were both visually and semantically distinct to cater to the requirements of memory assessment tasks. To avoid biases, this step was performed by two independent reviewers involved in the task design, and the agreement between the reviewers, calculated as Cohen’s kappa κ, was 0.87 (strong level of agreement; see McHugh 2012 for level of agreement for different values of κ). Cohen’s kappa coefficient is commonly used to measure inter-annotator reliability. It is a more robust measure than simple percentage agreement since it accounts for the possibility of agreement by chance. The remaining discrepancies were resolved by discussion.

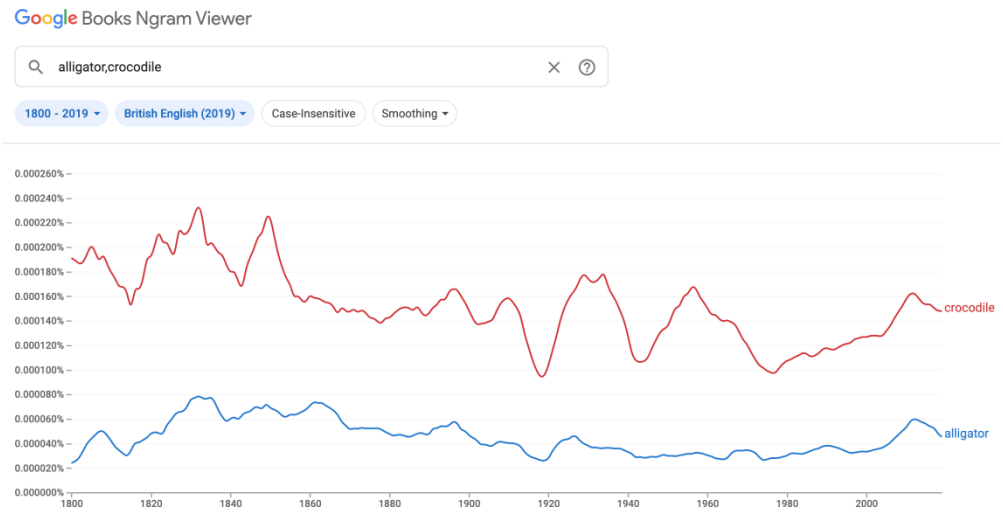


Figure 4.5. *Google n-gram frequencies of concept occurrence as a proxy for familiarity.*

4.2.4 The assembled image database

The discussed pipeline produced 1,402 distinct object images across 29 semantic categories. Examples of images from this dataset are shown in Figure 4.6 and the frequency of concepts across categories is plotted in Figure 4.7. The discussed pipeline highlighted the utility of the WordNet lexical database for automatically acquiring thousands of picturable concepts across semantic categories. Each image in the dataset has been manually confirmed to match with the WordNet definition of its corresponding concept synset. The concept names and their corresponding synsets in WordNet are made available in the appendix.

Figure 4.6. *Example images from the acquired dataset.*

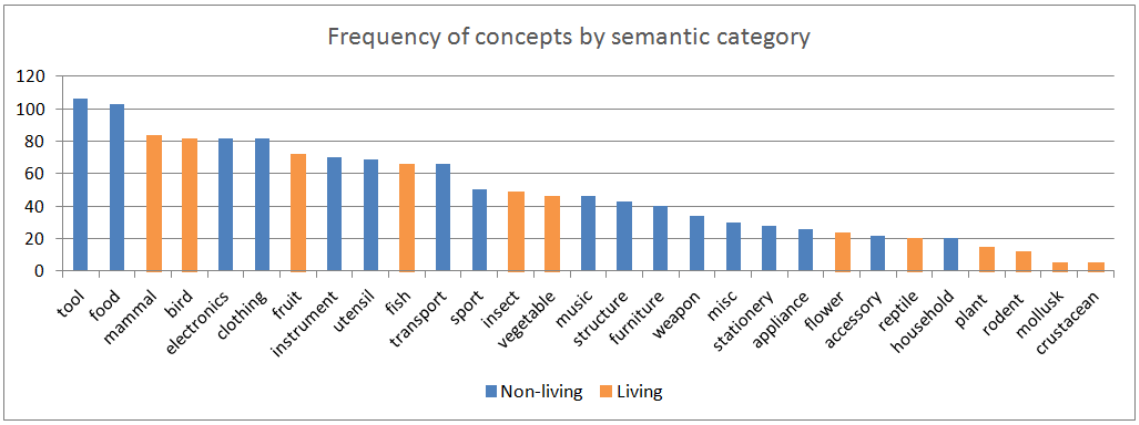


Figure 4.7. *Frequencies of concepts by semantic category in image database.*

4.3 Production of confusable object pairs using computational models

In this section, I implement 6 variants of the computational models of vision and semantics described in the previous chapter (by using them both alone and in different weighted combinations), to produce confusable object pairs from the object image dataset assembled in the previous section.

4.3.1 Linguistic similarity based confusable pair production

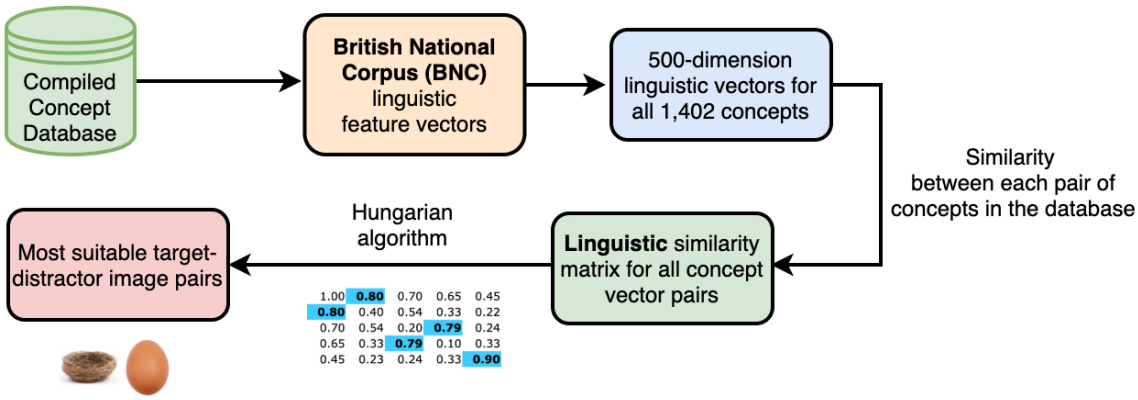


Figure 4.8. *Pipeline for generating concept pairs using the linguistic model.*

The pipeline for producing the dataset of confusable object pairs using the linguistic similarity model is shown in Figure 4.8. As discussed in section 3.2.1 in the previous chapter in detail, the linguistic feature vectors for all the 1,402 object concepts were

obtained from the Word2Vec word embeddings made available by Rei and Briscoe (2014). These were collated in a matrix as shown in Figure 3.3 in the previous chapter, where each of the 1,402 rows contained a 500-dimension word embedding representing that object concept per this model. By measuring the cosine similarity between all pairs of feature vectors in this matrix, a pair-wise linguistic similarity matrix with dimensions 1402 x 1402 was constructed. The value in position (*i, j*) in this matrix would indicate the similarity between the i^{th} and the j^{th} concepts per the linguistic model. Finally, the Munkres algorithm (section 3.2) was applied to this pairwise similarity matrix to obtain unique target-distractor object pairs parametrically varying in linguistic similarity.

4.3.2 Visual similarity based confusable pair production

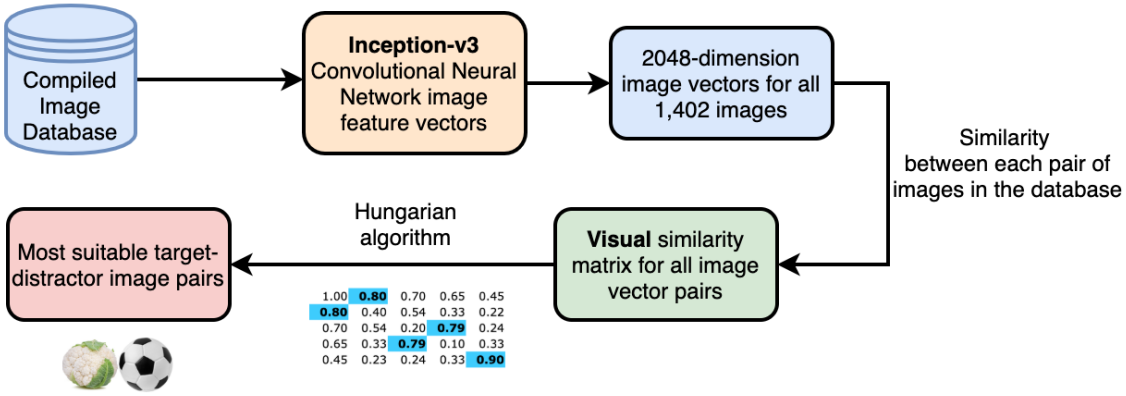


Figure 4.9. *Pipeline for generating concept pairs using the visual model.*

The pipeline for producing a dataset of confusable object pairs using the visual similarity model is shown in Figure 4.9. As discussed in section 3.2.2 in the previous chapter in detail, the visual feature vectors for all the 1,402 object images were obtained from the fully connected layer in the Inception-v3 deep-CNN model. Similar to the linguistic model, these feature vectors were collated in a matrix, where each of the

1,402 rows contained a 2048-dimension image vector representing that object image per this model. By measuring the cosine similarity between all pairs of feature vectors in this matrix, a pair-wise visual similarity matrix with dimensions 1402 x 1402 was constructed. The value in position (*i, j*) in this similarity matrix would indicate the similarity between the i^{th} and the j^{th} concepts per the visual model. Finally, the Munkres algorithm was applied to this pairwise similarity matrix to obtain unique target-distractor object pairs parametrically varying in visual similarity.

4.3.3 Taxonomic similarity based confusable pair production

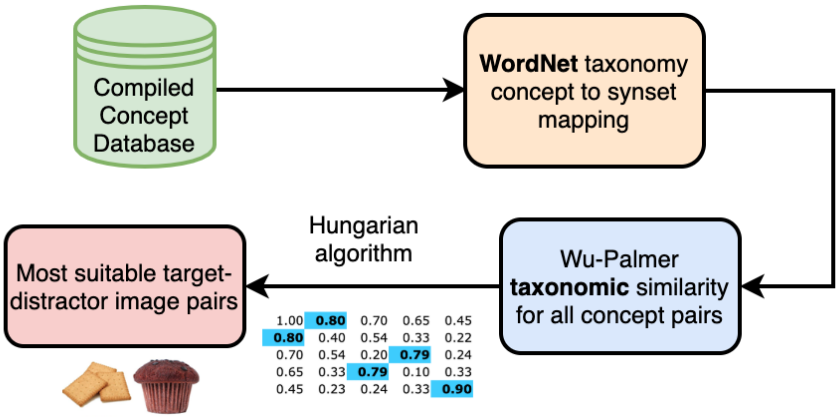


Figure 4.10. *Pipeline for generating concept pairs using the taxonomic model.*

The pipeline for producing the dataset of confusable object pairs using the taxonomic similarity model is shown in Figure 4.10. As discussed in section 3.2.3 in the previous chapter in detail, the WUP similarity metric for WordNet was used to obtain similarity between concept pairs in the WordNet taxonomy. As part of the image dataset obtained in section 4.1, WordNet concept definitions were used to identify specific synset labels for each image in the dataset (for instance, 'dog.n.01' for concept *dog*, 'crane.n.02' for concept *crane*, etc.; refer to the appendix for the complete list). These synsets were used to construct a matrix of pairwise WUP similarities between

concepts (with dimensions 1402 x 1402) such that the value in position (*i, j*) in the matrix refers to the taxonomic similarity between the i^{th} and the j^{th} concepts per the WUP similarity metric for the WordNet taxonomy.

4.3.4 Bimodal visuo-linguistic similarity based confusable pair production

As discussed in earlier chapters, while the unimodal models capture some aspects of human cognition (Landauer et al., 2007; Collel and Moens, 2016), humans do not acquire word meanings and the relationships between concepts through any one information source (Landau et al., 1998; Bornstein et al., 2004), but rather through a combination of these. It was also briefly discussed in Chapter 2 (section 2.3) how a number of studies have now shown that integrating information from multiple sources yields more complete concept representations, leading to a better estimation of the relationships between concepts – a problem central to this thesis. With enough evidence across studies of semantic representations from multiple sources performing better at capturing human knowledge of conceptual similarity (Bruni et al., 2014; Silberer and Lapata, 2014; Kiela and Bottou, 2014; Lazaridou and Baroni, 2015; Derby et al., 2018), I integrated the unimodal models to create new datasets of confusable object pairs.

First among these models was the visuo-linguistic model. Traditional machine learning based models require training the models from scratch, but these techniques work only when a large amount of data (such as a large number of images for each object concept) are available. The problem central to this thesis, however, is how to utilise the computational techniques for representing objects when only one image per object concept is available – an image that is to be used as a stimulus to judge participant response in a memory assessment task. A key consideration here, therefore, was to

use lightweight combination techniques that could support equitable representation from the constituent unimodal models in the resulting integrated model.

To facilitate this, a feature concatenation method introduced by Kiela and Bottou (2014) was used. This method worked as follows: the visual and linguistic feature spaces obtained in the previous sections were combined by concatenating the normalized, centred to the mean and component-wise scaled to unit variance visual and linguistic feature vectors for each of the 1,402 concepts. This involved appending the re-weighted linguistic feature vector for each object to the re-weighted visual feature vector for the corresponding object, where the re-weighting was performed as shown in equation 4.1. Since the image feature vectors are 2048-dimensional and the word embeddings are 500-dimensional, concatenation resulted in 2548-dimensional visuo-linguistic feature vectors. The construction of this vector space is elaborated next and is summarised in the pipeline presented in Figure 4.11.

Bimodal representation: Following previously reported procedures (Bruni et al., 2014; Kiela and Bottou, 2014), the following equation yields the visuo-linguistic feature vector for each concept:

$$\vec{v}_{\text{concept}} = \alpha \times \vec{v}_{\text{vis}} \,||\, (1 - \alpha) \times \vec{v}_{\text{ling}} \qquad (4.1)$$

where $\vec{v}_{ling}$ refers to the normalised and scaled linguistic feature vector, $\vec{v}_{vis}$ refers to the visual feature vector, || denotes the concatenation operator and α is the tuning parameter. The choice of the tuning parameter is explained next.

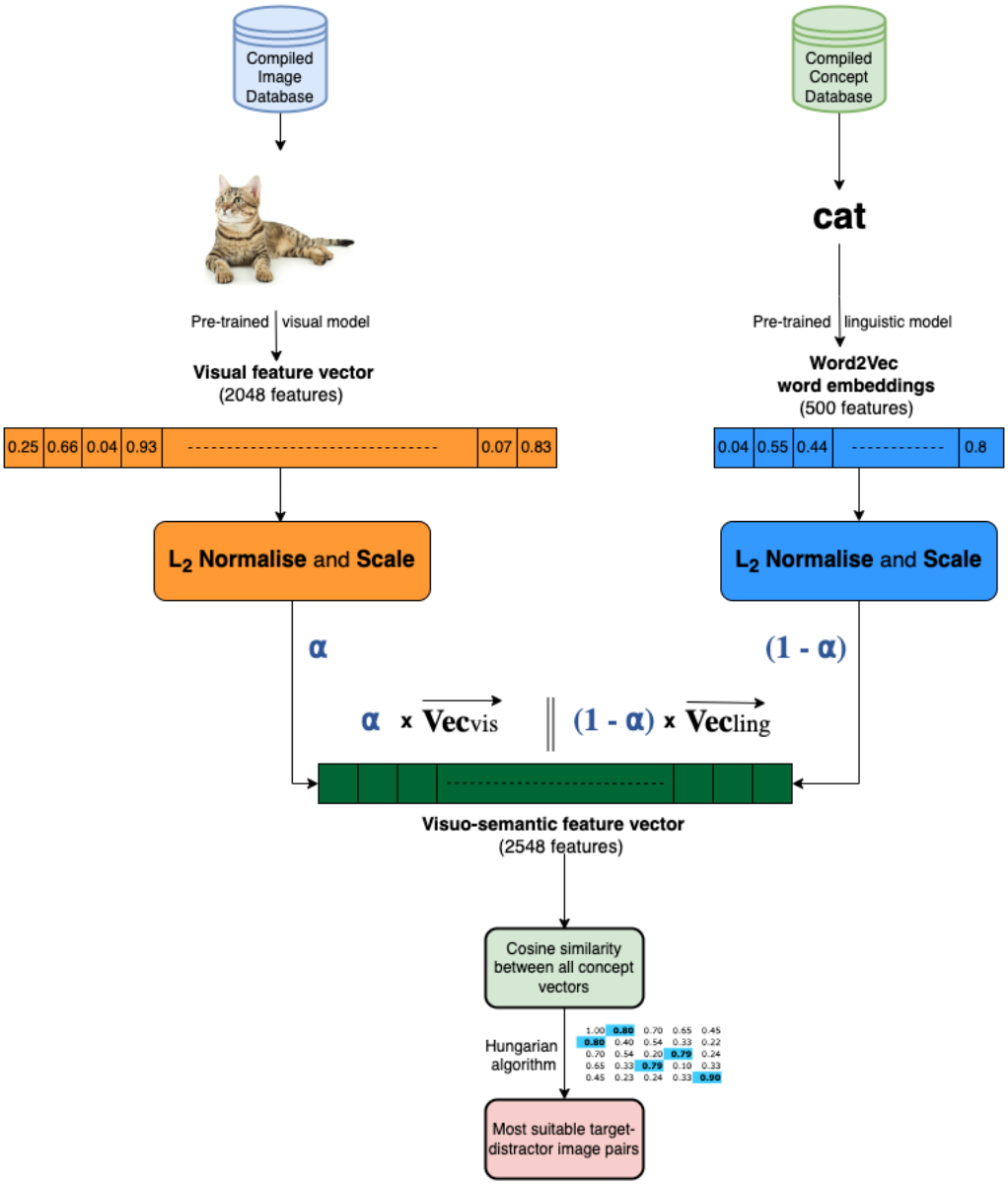
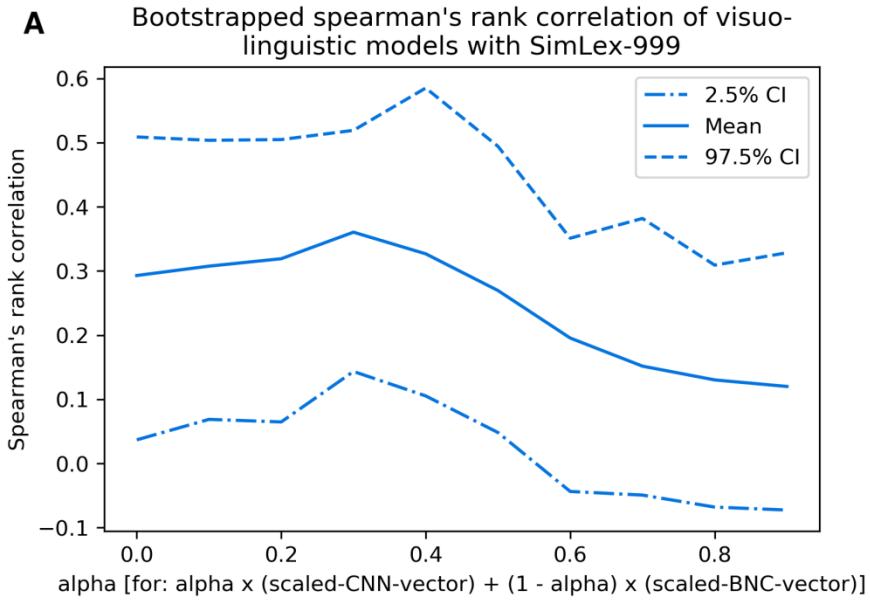


Figure 4.11. *Pipeline for generating concept pairs using the visuo-linguistic model. The visual feature vectors for the chosen images for each concept and the linguistic feature vectors for the corresponding concept words were first normalised and scaled to mean 0 and standard deviation 1. Following this, the linguistic feature vectors were concatenated onto the visual feature vectors. This combination or concatenation of vectors resulted in a combined vector with the number of dimensions equal to the sum of the individual dimensions of the two feature vectors. The pairwise cosine similarities between all new feature vectors were then calculated and passed through the Munkres algorithm to select the most suitable target-distractor image pairs.*

Tuning parameter: An important question to be addressed when concatenating information from two different sources is the weight to be assigned to each information source (α and 1 - α in equation 4.1) such that the hence weighted combination of these sources results in a vector space that maximises human understanding of

similarity between concepts. For this, traditionally the hyperparameter α is chosen such that it maximises the correlation of the resulting vector space with human-generated concept pair ratings. The human rated concept pair ratings are obtained from the standard evaluation datasets discussed in the previous chapter. The strategy to obtain the optimum value of the hyperparameter α can therefore be summarised as follows:

1. Substitute equation 4.1 with incremental values of α to obtain concept vectors for all 1,402 concepts for each value of α;
2. For each obtained vector space governed by a different value of α, obtain pairwise similarities for concept pairs common with human-rated evaluation benchmark datasets;
3. Measure the correlation between model-generated similarity indices and similarity indices from the human-rated datasets for the same pairs;
4. The value of α for which the correlation is maximum is the best value for the tuning hyperparameter.



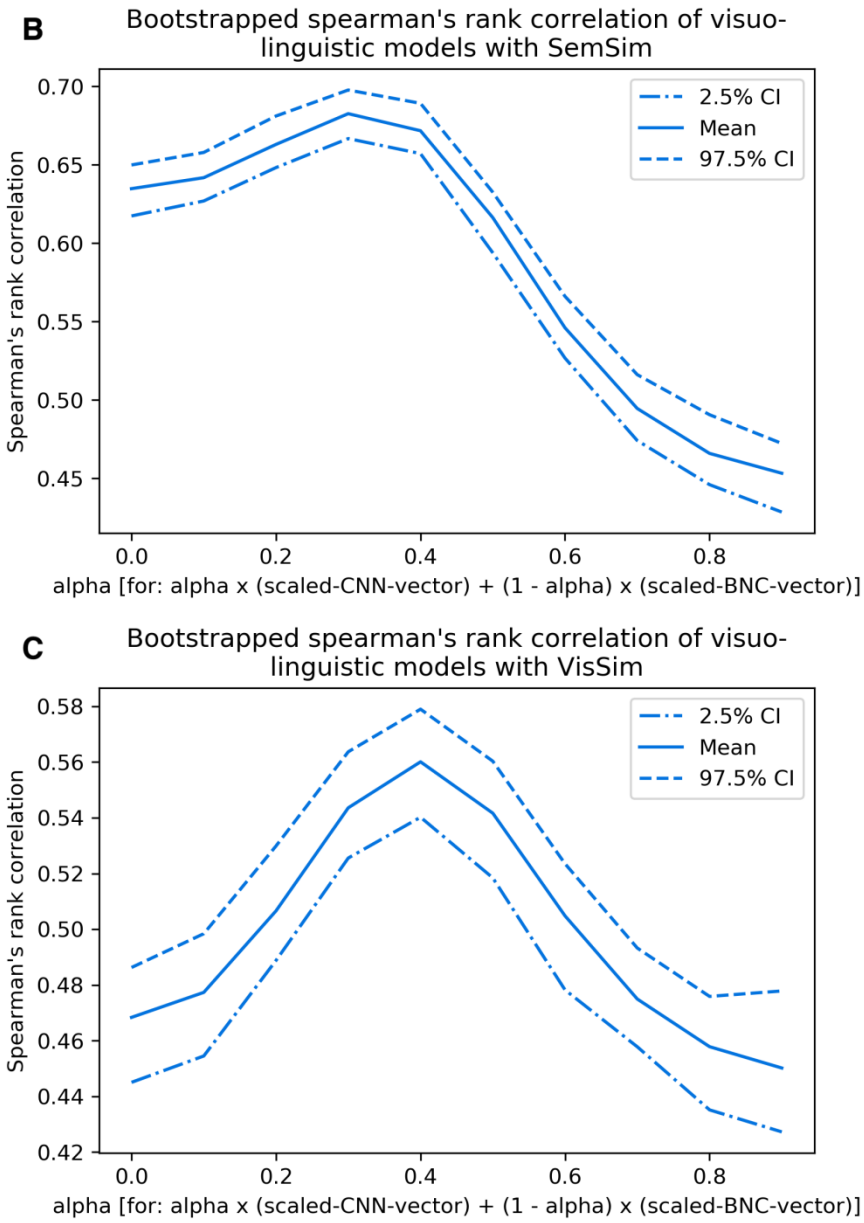


Figure 4.12. *Spearman's rank correlation of models produced using equation 4.1 with (A) SimLex-999, (B) SemSim, and (C) VisSim with confidence intervals for* $\alpha \in [0,1]$.

Following this scheme, values of α in increments of 0.1 with $\alpha \in [0,1]$ were substituted in equation 4.1 and fitted with common pairs from the SemSim and VisSim databases (Silberer and Lapata, 2014) and SimLex-999 (Hill et al., 2015) to obtain the best possible model. As is evident from Figure 4.12, the Spearman's rank correlation between pairwise cosine similarities between relevant[iii] visuo-linguistic concept vectors and

[iii] the term "relevant" applies to overlapping concept pairs between my database (of 1,402 concepts) and the evaluation databases used.

human ratings of similarities between such concept pairs was found to be maximum at α = 0.3 with SemSim and SimLex-999, and α = 0.4 with VisSim. α = 0.3 was chosen for the final visuo-linguistic concept representation, yielding equation 4.2 below:

$$\vec{v}_{\text{concept}} = 0.3 \times \vec{v}_{\text{vis}} \,||\, 0.7 \times \vec{v}_{\text{ling}} \qquad (4.2)$$

Similar to the earlier models, the feature vectors obtained from this model were collated in a matrix, where each of the 1,402 rows contained a 2548-dimension visuo-linguistic vector representing object per this model. By measuring the cosine similarity between all pairs of feature vectors in this matrix, a pairwise visuo-linguistic similarity matrix with dimensions 1402 x 1402 was constructed. The value in position (*i, j*) in this similarity matrix would indicate the similarity between the i^{th} and the j^{th} concepts per this model. Finally, the Munkres algorithm was applied to this pairwise similarity matrix to obtain unique target-distractor object pairs parametrically varying in visuo-linguistic similarity.

4.3.5 Bimodal visuo-taxonomic similarity based confusable pair production

In the absence of feature vectors from the WordNet database, the averaging technique from Kiela and Bottou (2014) was adopted to combine similarities from both visual taxonomic models. Using this technique, similarity values between corresponding concept pairs from the two models were combined using a tuning parameter as per equation 4.3. The technique is summarised in Figure 4.13.

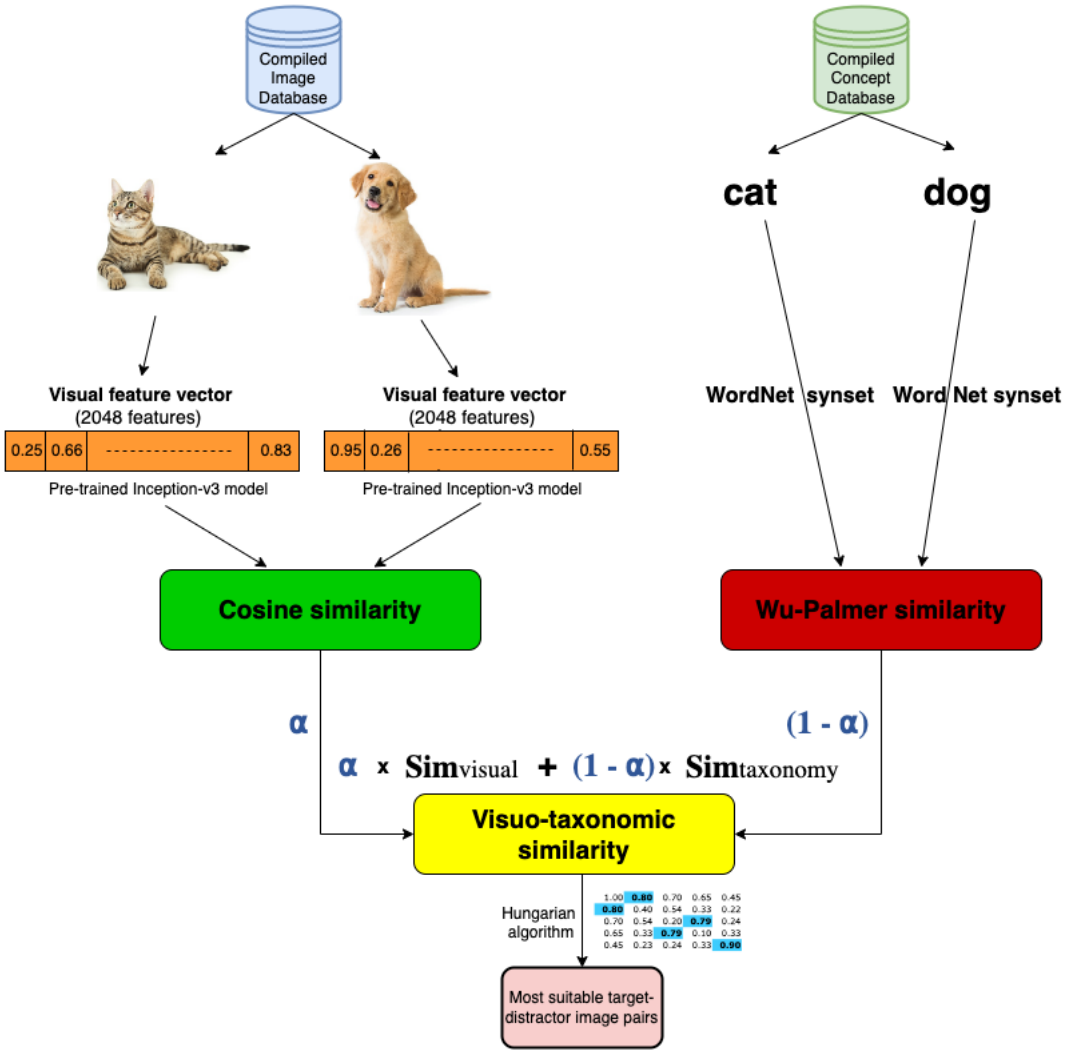
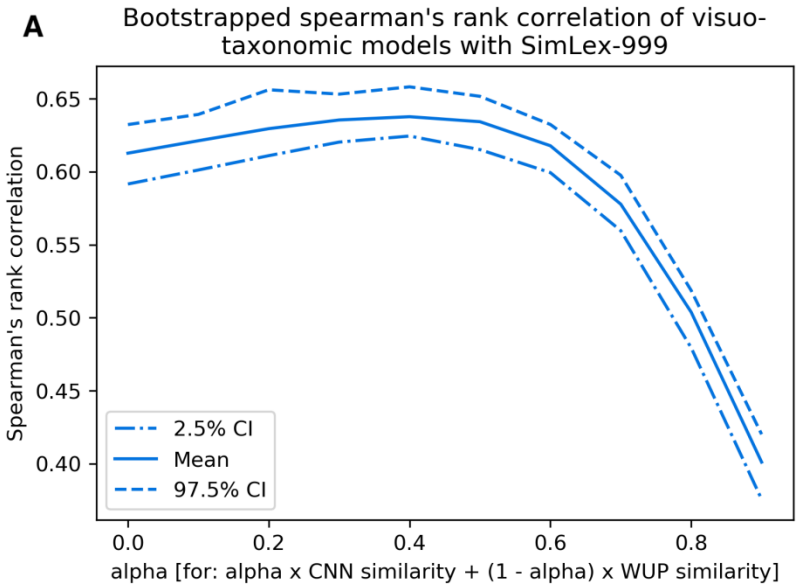


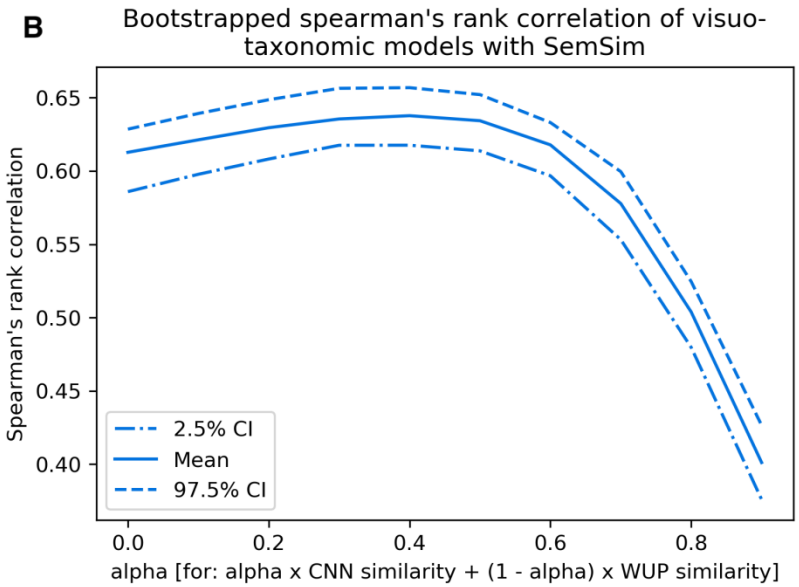
Figure 4.13. *Pipeline for generating concept pairs using the visuo-taxonomic model.*

The procedure for choosing the tuning parameter remained the same as in the visuo-linguistic model. Values of α in increments of 0.1 with $\alpha \in [0,1]$ were substituted in equation 4.3 below and fitted with common pairs from the SemSim and VisSim databases (Silberer and Lapata, 2014) and SimLex-999 (Hill et al., 2015) to obtain the best possible model. As is evident from Figure 4.14, the Spearman's rank correlation between pairwise visuo-taxonomic concept pair similarities and human ratings of similarities between such concept pairs was found to be maximum at α = 0.5 with SemSim, VisSim and SimLex-999. α = 0.5 was thus chosen for visuo-taxonomic similarity between concept pairs, stated by equation 4.3 below:

$$sim(A, B) = 0.5 \times sim(visA, visB) + 0.5 \times sim(taxoA, taxoB) \quad (4.3)$$

where sim(visA, visB) refers to the visual similarity between concepts A and B, sim(taxoA, taxoB) refers to the taxonomic similarity between concepts A and B and α is the tuning parameter.





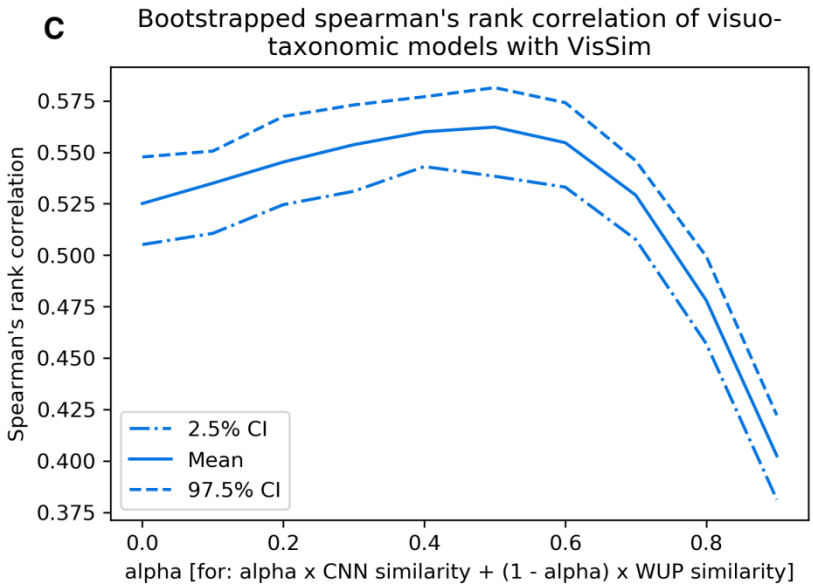


Figure 4.14. *Spearman's rank correlation of models produced using equation 4.3 with (A) SimLex-999, (B) SemSim, and (C) VisSim with confidence intervals for* $\alpha \in [0,1]$.

The same steps as in previous experiments were followed to obtain matched image pairs from this model as shown in the pipeline. Since the model directly generated pairwise similarities for all concept pairs, such similarities were then used to choose unique pairs using the Hungarian algorithm.

4.3.6 Retrofitted visuo-linguistic similarity based confusable pair production

While combining information from two sources has received a lot of attention in the last few years, usually involving visual and linguistic sources (Bruni et al., 2014; Silberer and Lapata, 2014; Kiela and Bottou, 2014; Lazaridou and Baroni, 2015), combining information from all the three sources presented in this thesis is rare. Having obtained bimodal models, I next present a novel attempt at combining information from all three sources. To facilitate this, I made use of a lightweight post-processing technique called retrofitting, which is discussed next.

Retrofitting: Faruqui et al. (2015) devised a graph-based learning technique for using lexical relational resources such as the WordNet database to obtain high quality

concept vectors from existing vectors. The idea behind the approach is to fine-tune the concept vectors using taxonomic information. Specifically, in this thesis, I have taken the visuo-linguistic vector space and updated each vector in this space to get closer and more similar to other members in its category obtained from the WordNet taxonomy. Intuitively, this ensures that the updated or "retrofitted" concept representation contains information from all three sources – visual, linguistic, and taxonomic. The key advantage of this approach is that instead of training a network from scratch, this approach can be used on pre-trained vectors. Since the vectors used in this thesis are pre-trained due to reasons explained in the previous chapter, this lightweight post-processing approach is useful here.

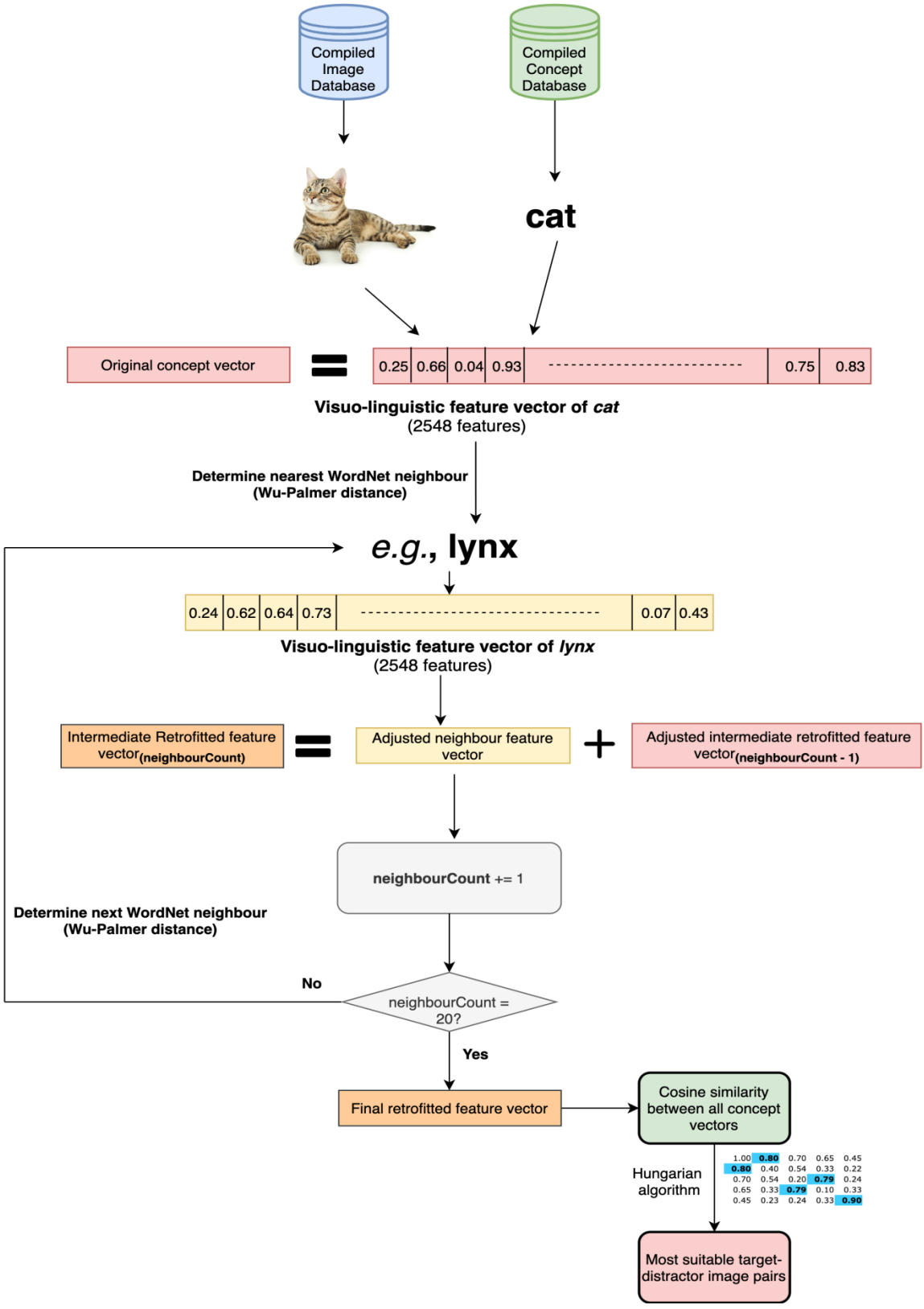


Figure 4.15. *Pipeline for generating concept pairs using the retrofitted visuo-linguistic model with 20 neighbours.*

Working of the retrofitting approach: Let $\hat{Q}$ be the collection of pre-trained vector representations $\hat{q}_i \in \mathbf{R}^d$, where *i* refers to each concept in the database and *d* is the number of dimensions of each vector. The objective is to learn a new, high quality semantic vector space Q from $\hat{Q}$ using a lexical database such that the vectors thus obtained are close to both $\hat{Q}$ and to adjacent vertices or neighbours in the lexical database. Faruqui et al. (2015) show that by using the online update shown in equation 4.4 below, such a vector space can be obtained:

$$q_i = \frac{\sum_{j:(i,j)\in E} \beta_{ij} q_j + \alpha_i \hat{q}_i}{\sum_{j:(i,j)\in E} \beta_{ij} + \alpha_i} \qquad (4.4)$$

where, q_i is the retrofitted vector representation of concept *i*, $\hat{q}_i$ is the original vector representation of concept *i*, (in this thesis, this is the visuo-linguistic representation of concept *i* from equation 4.2), *j* is an iterator that runs through the neighbours of concept *i* in the taxonomic database (in this thesis, the neighbours of a concept in the WordNet taxonomic database are determined using the WUP similarity metric) and consequentially q_i for each value of *j* is the vector representation of the j^{th} neighbour, E is the set of neighbours of concept *i*, and α and β are the values that control the relative strengths of association, where all α_i are set to 1 and β_{ij} to (degree)$^{-1}$, where degree is the number of neighbours chosen.

Implementation of the retrofitting approach: The Python implementation of the retrofitting approach is available at https://github.com/deveshbatra/confusable-pair-selection under "retrofit.py". The implementation has been summarized for intuition in Figure 4.15. The steps are explained briefly as follows:

1. For each concept, the original visuo-linguistic vector was obtained from equation 4.2. This is represented as $\hat{q}_i$ in equation 4.4.
2. The variable *j*, which represents the number of lexical neighbours in WordNet chosen to update or retrofit the vector, was chosen using the WUP similarity metric (the higher the WUP similarity metric for WordNet, the closer the neighbour). For example, the 5 closest neighbours to the concept cat in the WordNet database ranked by the WUP similarity metric are: *lynx*, *cheetah*, *jaguar*, *leopard,* and *lion*.
3. The original visuo-linguistic vector obtained in step 1 was updated using the corresponding neighbour vector (q_j in equation 4.4). For the purpose of this thesis, retrofitted vectors were obtained by varying the number of neighbours from 1 to 20, i.e., 20 different vector spaces (and hence 20 different vectors for each concept) were obtained by keeping the number of neighbours constant for all concepts in a single vector space. For instance, for the concept *cat*, the concept vector in the first vector space would be updated on *lynx* only; the second would be updated on *lynx* and *cheetah*; the third on *lynx*, *cheetah*, and *jaguar*, and so on.
4. Upon obtaining 20 different vectors for each concept (and thus obtaining 20 different vector spaces, each corresponding to the number of neighbours in WordNet used to update vectors in such spaces), the influence of the number of neighbours on the quality of retrofitted vectors was obtained using the evaluation benchmarks (SemSim and VisSim) used in earlier models. Specifically, for each vector space (1 to 20), Spearman's rank correlations between similarity values from retrofitted vector pairs and human-rated similarities for concept pairs from human-rated benchmark datasets were obtained and plotted as shown in Figure 4.16.

The pattern from this analysis can be interpreted as follows. The Spearman's rank correlation increases with both the VisSim and SemSim concept pairs until the number of neighbours reach 10, beyond which, increasing the number of neighbours has no significant effect on the vector's performance.

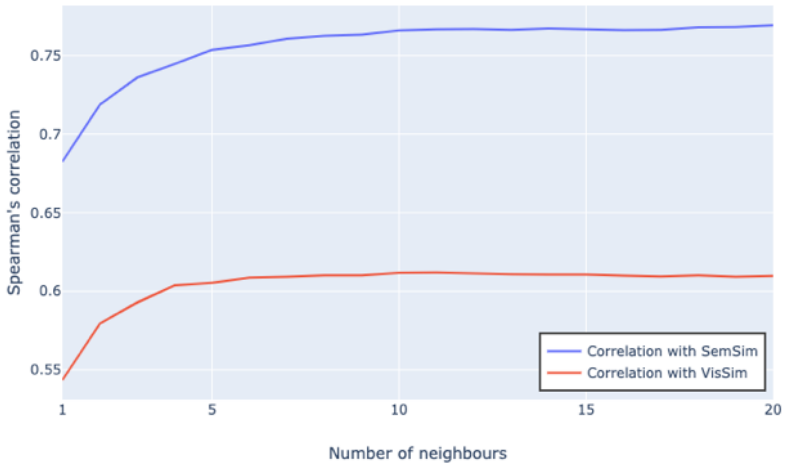


Figure 4.16. *Spearman's rank correlation of models produced using equation 4.3 with SemSim and VisSim for* $\alpha \in [0,1]$.

As a result, the model chosen for producing confusable object pairs after this evaluation was the one with number of neighbours = 10. The same steps as in previous experiments were followed to obtain matched image pairs from this model.

4.4 Analysis

All computational models implemented in this chapter were assessed on the two evaluation techniques discussed in section 3.3 in the previous chapter: 1) proportion of all object pairs produced by each model in which the constituent objects belonged to the same semantic category – the higher this metric, the better the model performance; and 2) Spearman's rank correlation with human judgements of object pair similarity. For the latter, SemSim, VisSim and SimLex-999 datasets were chosen for evaluation as discussed in the previous chapter. For reference, the interpretation values for Spearman's ρ commonly found in published literature are: 0.2

(small correlation), 0.35 (small-medium correlation), 0.5 (medium correlation), 0.65 (medium-large correlation) and 0.8 (large correlation). The higher the Spearman's rank correlation between the similarity ratings from a human-rated dataset and model-produced ratings, the more aligned a model's similarity ratings are with human understanding of object pair similarity. Since the aim of using these models is to produce parametrically varying object pairs, the models showing greater alignment with human-rated datasets would be better and preferred for further validation and use.

4.5 Results

The performance of the computational models on the evaluation metrics are summarised in Table 4.1 and plotted in Figures 4.17 and 4.18.

Table 4.1. *Summary of evaluation metrics for all computational models.*

Model	Percentage of pairs	SimLex-999 (ρ)	VisSim (ρ)	SemSim (ρ)
Visual	50.6%	0.12	0.45	0.45
Linguistic	48.8%	0.29	0.46	0.64
Taxonomic	65%	0.33	0.53	0.68
Visuo-linguistic	63.40%	0.36	0.54	0.68
Visuo-taxonomic	74.80%	0.36	0.57	0.69
Retrofitted Visuo-linguistic	76%	0.49	0.62	0.77

4.5.1 Model performance on semantic category membership

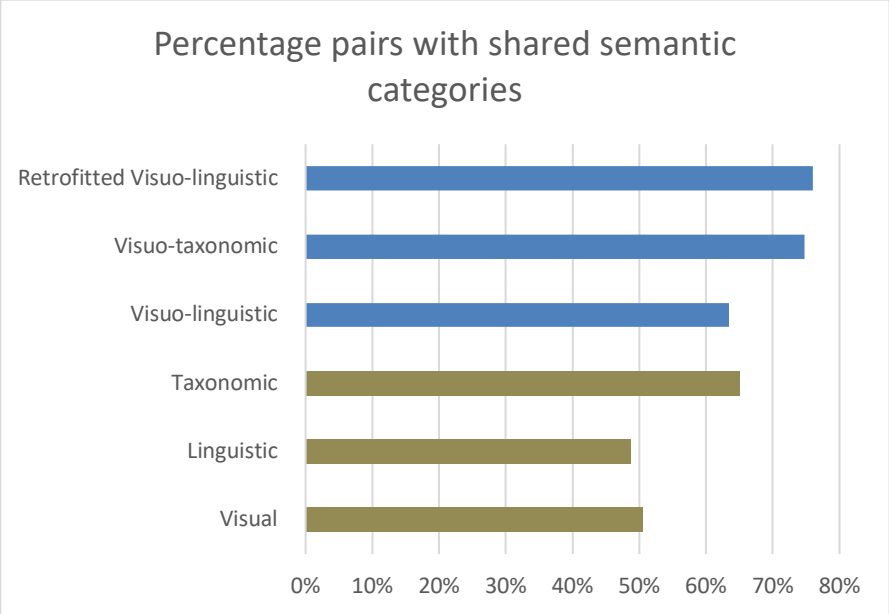


Figure 4.17. *Percentage pairs produced by each model with shared semantic categories.*

Multimodal models generally outperformed unimodal models in producing object pairs that comprised of objects belonging to the same semantic category, apart from the taxonomic model that marginally outperformed the visuo-linguistic model on this metric. While the linguistic-only model produced the lowest proportion of all object pairs in which the objects shared semantic categories (48.8%), 76% of all object pairs produced by the best performing retrofitted visuo-linguistic model on this metric shared semantic categories.

4.5.2 Model performance with human judgements of object pair similarity

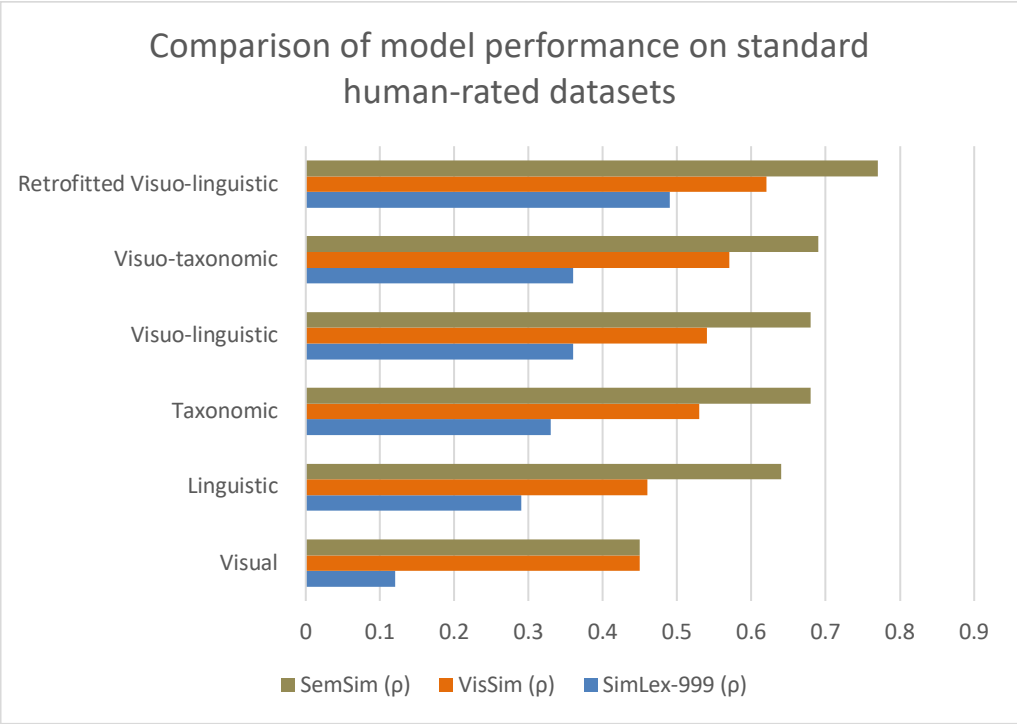


Figure 4.18. *Performance of computational models on standard evaluation datasets.*

As shown in Figure 4.18 and noted in Table 4.1, the multimodal models outperformed the unimodal models across the three standard human-rated datasets of object pair similarity. Across datasets, the retrofitted visuo-linguistic model emerged as the best performing model, followed by the visuo-taxonomic and visuo-linguistic models. Among the unimodal models, the taxonomic model outperformed the linguistic and the visual models in capturing the correlation with the object pair similarity rankings generated by the human-rated datasets.

4.6 Discussion

In this chapter, I showed how computational models of vision and semantics, as well as their combinations, could be used to produce datasets of confusable object pairs. Upon production of these datasets, the models, and the datasets they produced were examined against standard evaluation techniques. This evaluation was done to

facilitate a comparison among the models in order to find the best performing models, the confusable pairs produced by which could then be used for further evaluation and use in a memory assessment task later in this thesis. A side product of this analysis was an assessment of whether integrated computational models outperform the standalone models at estimating the human understanding of object pair similarity. This section elaborates on the findings obtained from the analysis conducted in this chapter.

4.6.1 Comparison among the unimodal models

Among the unimodal models, the taxonomic model outperformed both visual and linguistic models in performance across evaluation metrics. This result appeals to the intuition. Since the WordNet taxonomy is constructed purely on the basis of category-wise relationships between concepts, its performance on the semantic category membership metric is expected to be superior. Additionally, since objects belonging to the same superordinate category are likely to share both visual and semantic features, the taxonomic model is likely to capture a more complete estimate of overall similarity, thereby outperforming the other models on human-rated datasets of similarity.

On the evaluation of other unimodal models, the linguistic model outperforming the visual model on the SemSim dataset is not surprising either since the SemSim database has been rated to reflect the semantic similarity between object pairs. However, the former outperforming the latter on the VisSim dataset is an unexpected result since the VisSim dataset was rated on visual similarity between concept pairs, where ideally the visual model should have outperformed the linguistic model. This result can, however, be explained through the following theory: the existing datasets of human-rated similarity used in this chapter (including the VisSim dataset) could be biased towards word-based models of similarity since across these datasets, the

participants were shown words, and not images of concepts, for the purpose of rating similarities. As discussed in the previous chapters, such a setting has been criticised for allowing the participants to form their own mental models of objects during rating, thereby not reflecting reliable visual similarity ratings (see also Perfetti, 1998; Barsalou, 1999 and Glenberg and Kaschak, 2002 for support to this theory). Despite this constraint, these datasets were the best possible evaluation benchmarks at this stage of the thesis. A new dataset is developed in the next chapter that addresses the absence of a picture-based human-rated dataset in the field.

For a qualitative analysis of the capabilities of the respective unimodal models in object pair production, typical examples of object pairs obtained from these models are presented in Figure 4.19. As expected, the visually and semantically similar (*left*) and dissimilar (*right*) pairs reflect the capabilities of each model. For instance, since the linguistic model is based on "relatedness" or "associativity" between objects, it produces pairs such as <*egg, nest*> (Figure 4.19A) that are related and appear in similar contexts but are not similar and hence not confusable from a visual standpoint. Similarly, the visual similarity based deep-CNN model can be criticised for producing pairs such as <*snake, rope*> (Figure 4.19B) with constituent objects belonging to different semantic categories and thus not being semantically grounded. Finally, the WordNet-based WUP similarity metric can also be criticised for producing pairs such as <*biscuit, muffin*> (Figure 4.19C) that are categorically similar but not grounded in visual similarity.

Figure 4.19A. *Examples of matched pairs produced by the linguistic model. Both pairs are semantically related (co-occur), but only <bicycle, motorbike> is visually similar, whereas <nest, egg> is visually dissimilar.*

Figure 4.19B. *Examples of matched pairs produced by the visual model. While <cow, horse> pair (left) is both visually and semantically similar, <snake, rope> pair (right) is only visually similar, but semantically unrelated.*

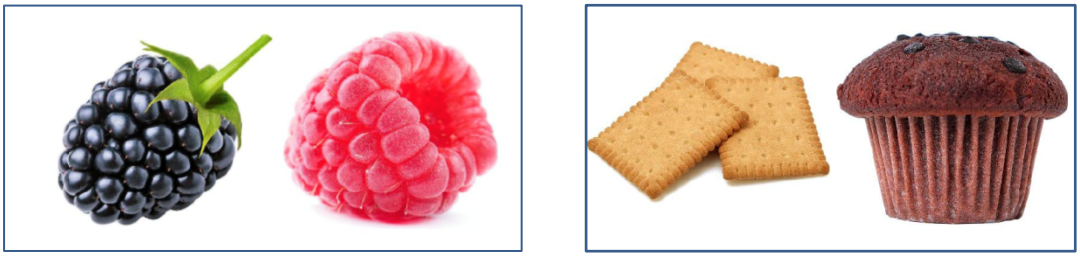


Figure 4.19C. *Examples of matched pairs produced by the taxonomic model. Both pairs are semantically similar, but only <blackberry, raspberry> is visually similar, whereas <biscuit, muffin> is visually dissimilar.*

4.6.2 Bimodal vs. unimodal models

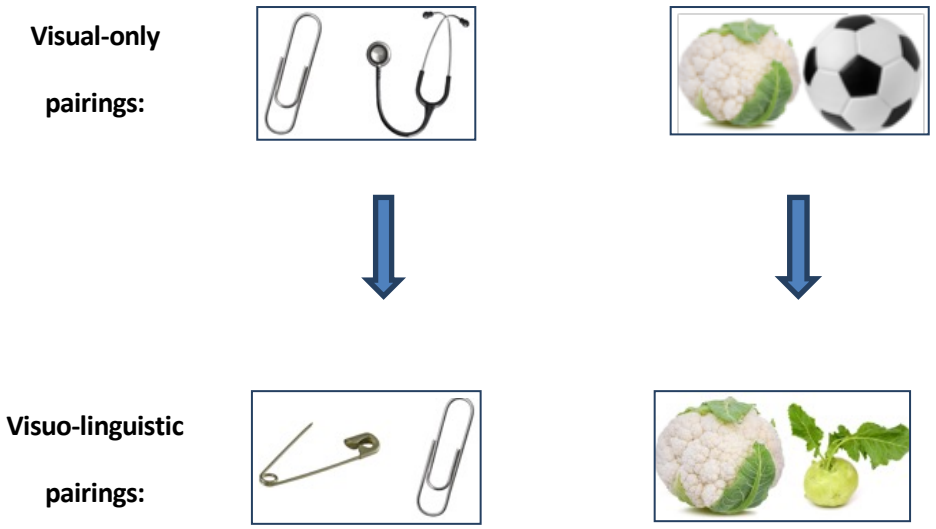


Figure 4.20. *Qualitative example of improvement in deep-CNN model performance upon grounding in semantic information.*

4.6.3 Wordnet-retrofitted visuo-linguistic model

The wordnet-retrofitted visuo-linguistic model outperformed all other models across evaluation measures. Unlike prior work, in this method, visual, linguistic, and taxonomic computational models were combined using a lightweight post-processing technique first introduced by Faruqui et al. (2015). As mentioned previously in this thesis, a crucial problem that psychiatrists face is forming high quality confusable object pairs using single images of object concepts in a logistically efficient manner. Since training large neural networks for this purpose would require multiple images of each concept in addition to large number of resources and time, lightweight post-processing methods such as retrofitting provide a middle ground by facilitating production of high-quality object representation vectors without the need to train machine learning models from scratch. This way, the method adds to the arsenal of computational techniques such as one-shot learning (used with other models

implemented in this chapter) that facilitate easy access to the knowledge of vast neural networks for use in applications such as those central to this thesis.

The work presented by Faruqui and colleagues, while ground-breaking, did not however assess the impact of increasing the number of lexical neighbours on the retrofitted vectors. The work on retrofitting presented in this chapter, therefore, extends the work done by Faruqui and colleagues by showing how the performance of retrofitted feature vectors changes with an increase in the number of lexical neighbours used to update the original feature vectors. Results showed that increasing the number of neighbours used to retrofit vectors increases the performance of the retrofitted model up to a certain level (#neighbours≈10), after which there are no significant further improvements. This result appeals to the intuition since exposing the vectors to their immediate lexical neighbours (for instance, *lynx, cheetah, jaguar, leopard, lion* for the concept *cat*) should indeed make the vector representation closer to human understanding. However, after the number of neighbours reaches 10, the neighbours used to update or "retrofit" concept vectors may now not be similar enough (for instance, other non-feline MAMMALS for updating the vector of concept *cat* such as *horse*), leading to the saturation of vector performance. This result was useful in choosing a suitable retrofitted model in this chapter.

4.7 Conclusion

This chapter presented the methodology followed in this thesis to generate datasets of confusable object pairs using computational models of vision and semantics, as well as their weighted combinations. Extending previous work done in the field, creation of new multimodal models combining information from standalone resources was also demonstrated in this chapter. Standard evaluation techniques, especially, Spearman's rank correlation between model-generated and human-generated similarity ratings

were used to compare model performances. The superior Spearman's rank correlations of multimodal models across human-rated datasets suggest that the similarity ratings produced by these models align closer to those produced by humans in comparison to the unimodal models. This indicates that the multimodal models are better at approximating human understanding of object pair similarity and should be preferred for producing parametrically varying confusable object pairs for use in memory assessment tasks. Additional work, however, is needed to validate each pair produced by such models before using them in memory assessment tasks.

Chapter 5

ImSim1498: a dataset of human-rated similarity values for computationally matched object pairs

5.1 Introduction

In the previous chapter, I showed how computational models of vision and semantics, as well as their weighted combinations, can be used to produce datasets of perceptually confusable object pairs. These models were individually rated against existing human-rated datasets to evaluate their efficacy in estimating human knowledge of concept pair similarity. While this approach is helpful in facilitating comparison and ranking of computational models, it does not provide a human-validated similarity rating for each individual pair produced by these models needed for further assessment of the usability of these pairs in cognitive assessment tasks. Moreover, since the similarity ratings in the evaluation datasets were obtained on word pairs (and not image pairs), they limit the precision of such datasets in capturing the "perceived" similarities between object images such as those central to this thesis, thereby not accurately recording the ability of the computational models in capturing this perceived similarity.

In light of the above requirements, this chapter presents the development of a new similarity rating questionnaire, aimed at capturing perceived similarity ratings of the computationally matched object pairs produced in the previous chapter. Before developing this questionnaire, designs of existing similarity rating questionnaires first introduced in Chapter 3 were reviewed. The best practices learned from these were incorporated in the production of the new questionnaire. The newly designed questionnaire was then deployed on a popular web-based crowdsourcing platform, the Amazon mechanical Turk, to obtain the required similarity ratings. A large sample of

US and UK-based participants was recruited through the platform, who remotely attempted the questionnaire under this Amazon mechanical Turk study.

Following obtaining and cleaning the acquired ratings from the study in a post-processing activity, a key second aim of the chapter was to demonstrate the reliability of the thus acquired dataset. This was done by testing how closely aligned the ratings from different raters for each image pair were, which is a standard metric for measuring the reliability of these datasets (Padó et al., 2007; Reisinger and Mooney, 2010; Silberer and Lapata, 2014). The expectation was that this interrater agreement across all image pairs would compare favourably with that obtained in the past datasets. The third aim of this chapter was to present an evaluation of the computational models used in this thesis against this new dataset of perceived object pair similarity ratings. Similar to the previous chapter, this was done by calculating how the similarity ratings from the computational models correlate with the similarity ratings from the new dataset, thereby estimating how well these models capture human understanding of perceptual similarity between object image pairs. The obtained correlations were then compared with those observed in word-based similarity rating datasets used in the previous chapter. The expectation was that these correlations would be different from those observed with the word-based similarity rating datasets used in the previous chapter, with the models capturing perceived similarities over word similarities ranked higher with the new dataset. Following this, the final aim of this chapter was to present an evaluation of the usability of the computationally generated object pair similarity ratings in downstream tasks (here, memory assessment) by comparing them to the (cleaned) average ratings obtained from the human rating task using methods employed for such an assessment in past literature. This assessment assists in deciding whether computationally generated ratings can be used for stimulus-level analysis of participant response in memory assessment tasks, or the cleaned human ratings would be a better choice for such an analysis.

5.2 Similarity rating questionnaire development

5.2.1 Key traits of existing similarity rating questionnaires

The studies previously designed to produce the widely accepted similarity rating datasets (Rubenstein and Goodenough, 1965; Miller and Charles, 1991; Finkelstein et al., 2001; Bruni et al., 2013; Silberer and Lapata, 2014; Hill et al., 2015) mentioned in the previous chapters were reviewed along with their critiques (Budanitsky and Hirst, 2006; Hill et al., 2015) to identify the essential traits to be taken into consideration for the development of a new study to obtain similar ratings. These traits are summarised below and have assisted in designing the questionnaire central to this chapter. Summary statistics of these datasets are presented in Table 5.1.

Table 5.1. *Summary statistics of the reviewed datasets.*

Dataset name	Authors (year)	Number of concept pairs rated	Number of ratings per pair
Rubenstein-Goodenough dataset	Rubenstein and Goodenough (1965)	65	51
Miller-Charles dataset	Miller and Charles (1991)	30	38
WordSim-353 dataset	Finkelstein et al. (2001)	353	13-16
MEN dataset	Bruni et al. (2013)	3000	50
VisSim and SemSim datasets	Silberer and Lapata (2014)	7,576	5
SimLex-999 dataset	Hill et al. (2015)	999	50

1. Similarity vs. association: The reviewed studies chose to rate either similarity or association between concepts based on the purpose central to their study. In either case, the importance of clarifying the difference between similarity and association to the participants before presenting them the rating exercise was key. Hill et al. (2015), for instance, used the examples of concepts *car* and *petrol* to illustrate relatedness or association, and that of *car* and *bike* or *cup* and *mug* to illustrate similarity to their participants before rating. Following this, they specified that the study seeks to rate similarity over association. Similarly, Silberer and Lapata (2014), who were seeking to separately obtain visual and semantic ratings for their word pairs, clearly specified that visual similarity refers to "how similar the objects look" whereas semantic similarity refers to "how similar the objects are in meaning" (see Silberer 2015, p.154). Following this, they also showed example ratings for word pairs for both the activities to clarify how rating should be done. However, the rating exercises conducted before these were less descriptive about the differences between similarity and association (Finkelstein et al., 2001; Bruni et al., 2013), which was shown to impact the reliability of their ratings (see Hill et al., 2015 for a review). In sum, a new rating study should disambiguate similarity from relatedness (or association) for the participants before they begin rating and ensure that they understand which of these two the study expects them to rate. For this purpose, definitions and examples can be used.

2. Contextual vs. context-free rating: Another key design decision highlighted by the reviewed studies was whether to gather contextual or context-free ratings. Here, context refers to the sense of the word. For instance, in the Rubenstein-Goodenough and Miller-Charles studies, subjects were asked for similarity ratings between *glass* and *tumbler* as well as between *glass* and *jewel*. Here, the word *glass* could refer to either the KITCHEN UTENSIL or the MATERIAL. It was expected that the subjects would either use the dominant sense of the target word

or assume the sense of the word based on the word in comparison. This context-free design paradigm was also used in Finkelstein et al. (2001), Bruni et al. (2013) and Hill et al. (2015). On the other hand, Silberer and Lapata (2014) used contextual ratings by mentioning the category of the word in parenthesis, e.g., *mouse_*(animal) and *mouse_*(computer). The critiques of context-free ratings argue that such ratings introduce a higher degree of subjectivity in the design process, since the word sense could vary from one rater's opinion to another's, thereby reducing the inter-rater agreement on the similarity rating of such word pairs (Yong and Foo, 1999; Budanitsky and Hirst, 2006). Budanitsky and Hirst argue that context-free ratings may be obtained when finding associations but not similarities between words. When finding similarities between concepts, the interest lies in the relationship between senses, while words are mere surrogates for those senses. Thus, when obtaining similarity, contextual ratings would outweigh context-free ratings and should therefore be preferred.

3. Consistency and reliability: Due to a high degree of subjectivity in similarity ratings, obtaining consistent and reliable ratings is a challenge in such studies. To overcome this challenge, the reviewed studies have relied on two key design decisions: 1) training the raters before rating, and 2) introducing tests or checkpoints in the task to allow for easy identification of data from participants who did not understand the task well or did not pay attention while performing the task. The former includes showing examples of the setup presented in the real task to clarify to the raters what is expected of them. Participant responses on checkpoints or test questions, on the other hand, have been used in cleaning the obtained responses using suitable statistical techniques before compiling final similarity ratings (especially in Hill et al., 2015; Silberer and Lapata, 2014). Both these measures have been shown to increase the reliability of the ratings compiled in these studies.

4. Number of annotators: As noted in Table 5.1, most of the reviewed studies employed 30 or more annotators in their rating exercise, with the exception of Finkelstein et al. (2001) and Silberer and Lapata (2014). Hill et al. (2015) argue that a low number of annotators (e.g., 5 for VisSim and SemSim and 13-16 for MEN) inflates the inter-rater agreement due to low number of comparisons made to calculate this metric. Since the inter-rater agreement is a representation of the reliability of a dataset, this in turn inflates the perceived reliability of a dataset. Additionally, a low number of annotators is likely to increase the bias in the ratings. As a thumb rule, therefore, it can be concluded that at least 30 participants should be employed for such rating exercises to avoid the noted shortcomings.

5. Tool used to rate similarity: Among the reviewed studies, the two commonly used tools to measure similarity were Likert scales (WordSim-353; SimLex-999; VisSim and SemSim) and shuffled decks (Rubenstein and Goodenough database). Likert scales allowed the annotators to mark the similarity on a numeric range (e.g., 1-7), where the lowest value typically represented “very dissimilar” and the highest value, “very similar”. On the other hand, in the shuffled decks method, the participants were asked to arrange a deck of cards (each of which had a unique noun pair printed) to rank pairs in decreasing order of similarity. The shuffled decks method provided participants with a global context of similarity since they could compare their ranking with other pairs in the deck. However, while this method provides ranks of pairs based on similarity, it does not provide the magnitude of similarity for each pair. Additionally, as the number of pairs to rate increases, this method is likely to become harder for the annotators to rate similarities (e.g., a deck of 1500 pieces of paper would be much harder to arrange). The Likert scale addresses both these shortcomings since it can be

used to rate the similarity values and is independent of the scale of concept pairs to be rated.

6. Grouping of contextually related pairs: As identified, unlike the shuffled decks method, Likert scale method does not provide an easy way for the annotators to compare their similarity ratings with other pairs in the study. To address this shortcoming, Hill et al. (2015) grouped the word pairs belonging to the same parts-of-speech in their study. Therefore, by grouping contextually related pairs, they facilitated a comparison among pairs for their annotators, which in turn helped in achieving more accurate ratings.

5.2.2 New similarity rating task format

The new similarity rating task (presented in full in section 5.3.3) was built upon the identified traits. Following the lead of the more recent studies (Bruni et al., 2013; Silberer and Lapata, 2014; Hill et al., 2015), the task was hosted on the Amazon mechanical Turk platform (https://www.mturk.com), which is a crowdsourcing website that helps researchers in collecting and annotating data for answering scientific questions that require input from human participants. The platform allows for the recruitment of larger and more demographically diverse samples of participants than those that can be recruited in lab-based settings, thereby yielding ratings that are more representative and reliable (Buhrmester et al., 2016; Casler et al., 2013). Once the task was hosted on the platform, select members of the platform willing to participate in the study first attempted a qualifying test, which consisted of instructions for the task and two qualifying questions. The instructions and the qualifying questions ensured that the participants understood what the task expected of them. Only upon passing the qualifying test could the participants attempt the main similarity rating questionnaire that followed. The main questionnaire consisted of the to-be-rated image pairs grouped by semantic categories. This grouping ensured that the participants

could compare their ratings among contextually similar image pairs at the time of rating. To mark the ratings for the image pairs, the participants used a fine-grained version of the Likert scale (on a scale of 100), also known as the visual analogue scale. The detailed design of the questionnaire is discussed in section 5.3.3, where the specific design choices made to accommodate the identified key traits in the questionnaire are elaborated.

5.3 Materials and methods

5.3.1 Participants

Before the participants could attempt the qualifying test needed to proceed to the main questionnaire, they had to satisfy three requirements. The participants needed 1) to be either UK or US residents; 2) have more than 100 previous assignments approved on the platform and 3) have a 97% or higher approval rate on the platform. The latter qualifications ensured that the participants were experienced with the platform and had taken their work seriously in the past. The requirement for the participant to be a resident of the United States or United Kingdom was because the instruction set for the study was in the English language and it was vital for the success of the study that the participants understood exactly what was expected of them. However, since the main questionnaire is only based on pictures and not words, there was no requirement for the participant to be native English speakers. There was, however, a requirement for the participants to acknowledge that they had normal or corrected to normal vision.

Upon meeting the requirements listed above, 797 Amazon mechanical Turk members attempted the qualifying test. 243 of these did not pass the qualifying test. In all, 554 participants (322 females) attempted the main questionnaire. Their descriptive data is shown in Table 5.2. The study protocol was ethically approved (University of Oxford Medical Sciences Inter-Divisional Research Ethics Committee: R60073/RE002). E-

consent was required upon study sign-up. All procedures were in accordance with the Helsinki declaration.

Table 5.2. *Demographic characteristics of participants from GameChanger round 2 selected for analysis.*

Demographics	
Overall *n*	554
Number of participants by age group	
18-30	215
30-40	189
40-50	83
50-60	44
60-70	20
70-80	3
80+	0
Number of participants by biological sex	
Male	229 (41%)
Female	322 (58%)
Non-binary	3 (<1%)
Number of participants by country	
US	527
UK	27

5.3.2 Stimuli

The image pairs produced by the three top performing computational models from the previous chapter (i.e., the visuo-linguistic model, the visuo-taxonomic model, and the retrofitted model) were rated for similarity in the presented study. These top three computational models produced 1,498 unique image pairs. Due to logistical constraints (e.g., size of the questionnaire) and study costs, pairs produced by all the six models could not be rated. Examples of these image pairs are shown in Figure 5.5.

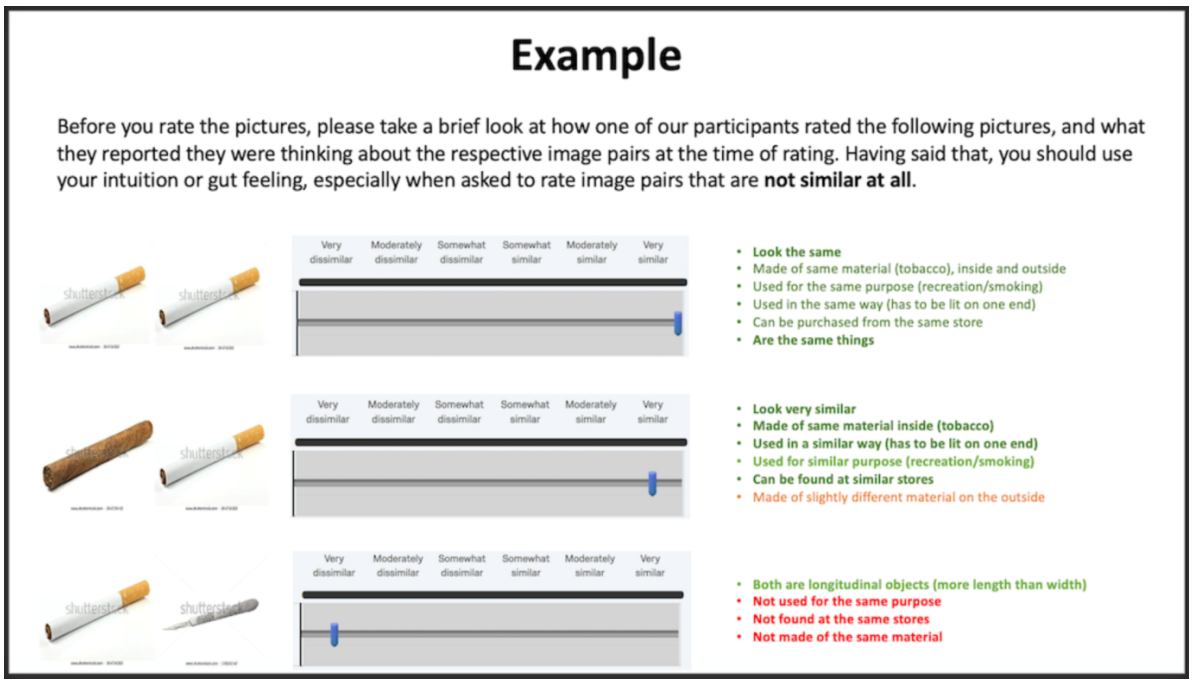
5.3.3 Task design

5.3.3.1 Qualifying test

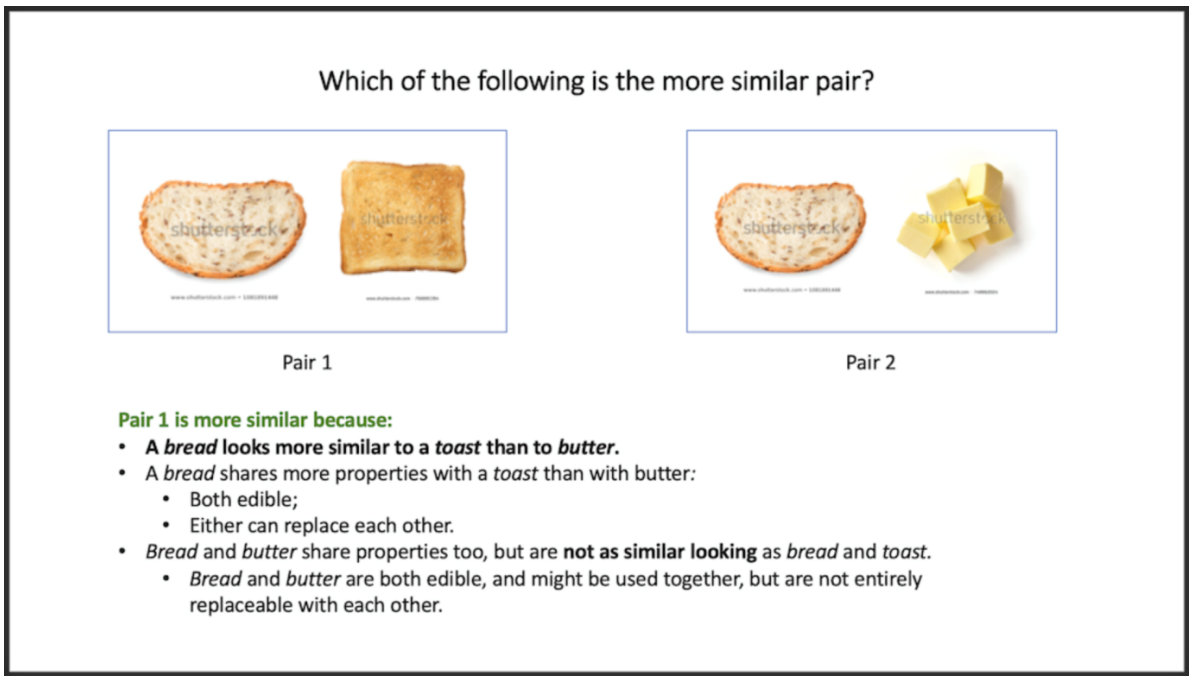
As noted above, the qualifying test consisted of two components: 1) training and 2) qualifying questions. Two key issues addressed in the training section were disambiguation between similarity and association, and disambiguation between visual-only similarity and overall perceived similarity between images. First, the participants were explained the requirements of the task with the following text:

"As humans, when we look at an object, we are able to rapidly process its meaning. This also allows us to understand how similar two objects are. For example, we understand that a tiger is more similar to a lion than to a sheep… In the present task, you will be shown a group of object pairs and asked to rate how meaningfully similar you think the two objects in the pair are on a non-discrete scale… For example, if a lion and a tiger are shown as a pair and you believe lion is meaningfully very similar to a tiger (since both are wild animals and hunt other smaller animals, etc.) or a snake and a rope are shown, and you believe they are not at all similar (since both are different categories of objects) you indicate that using a slider on a scale."

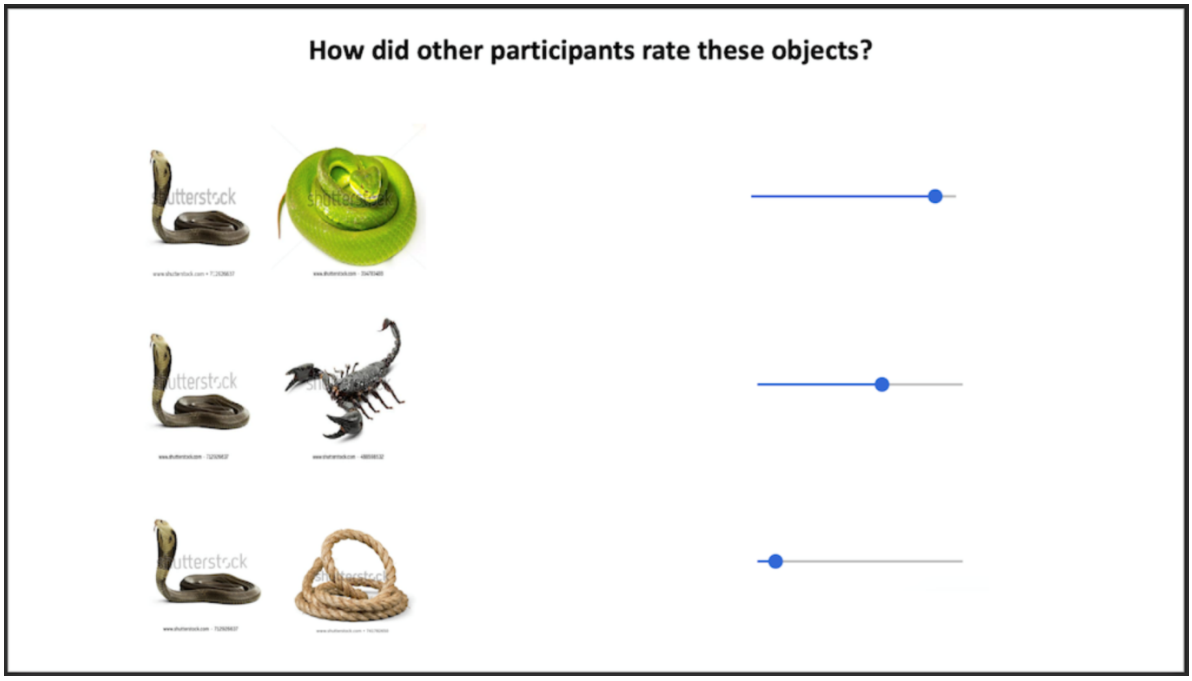
Following this, as shown in Figure 5.1A, the examples of the "*cigarette—cigarette*", "*cigar—cigarette*" and "*cigarette—scalpel*" pairs were used to show respectively that a) only two indistinguishable concepts obtain the highest value on the similarity scale, b) visually similar items that share function and meaning are rated higher on the scale and c) items sharing visual attributes only (attributes <*elongated*> and <*can be held between fingers*> in this case) with no commonality in function or meaning are rated very low in similarity. As shown in Figure 5.1B, the examples of "*bread—toast*" and "*bread—butter*" were used to exhibit the clear difference between similarity and association. In both these graphics, explanations of decisions behind ratings in the form of easily readable bullet points were given. In the last graphic (Figure 5.1C), without providing any explanation (in order to appeal to the intuition of the participant), it was shown how the main questionnaire would look like and how a typical participant would rate these pairs. At the end, participants were required to sign the declaration of having read the instructions carefully before submitting the qualifying test.



(A)



(B)



(C)

Figure 5.1. *Examples used to train participants.*

Following the training section, two test image pairs were presented to the participants as shown in Figure 5.2. Participants rating the pair "*hippopotamus—walrus*" more

similar than dissimilar (>=5 on a scale of 10) and rating the pair "*walrus—sack*" more dissimilar than similar (<5 on a scale of 10) qualified for the main questionnaire. Due to the unavailability of a slider-like or Likert scale-like tool in the Amazon Web Services framework for designing qualifying tests, radio buttons with values ranging from 1 (Not at all similar) to 10 (Strongly similar) had to be used.

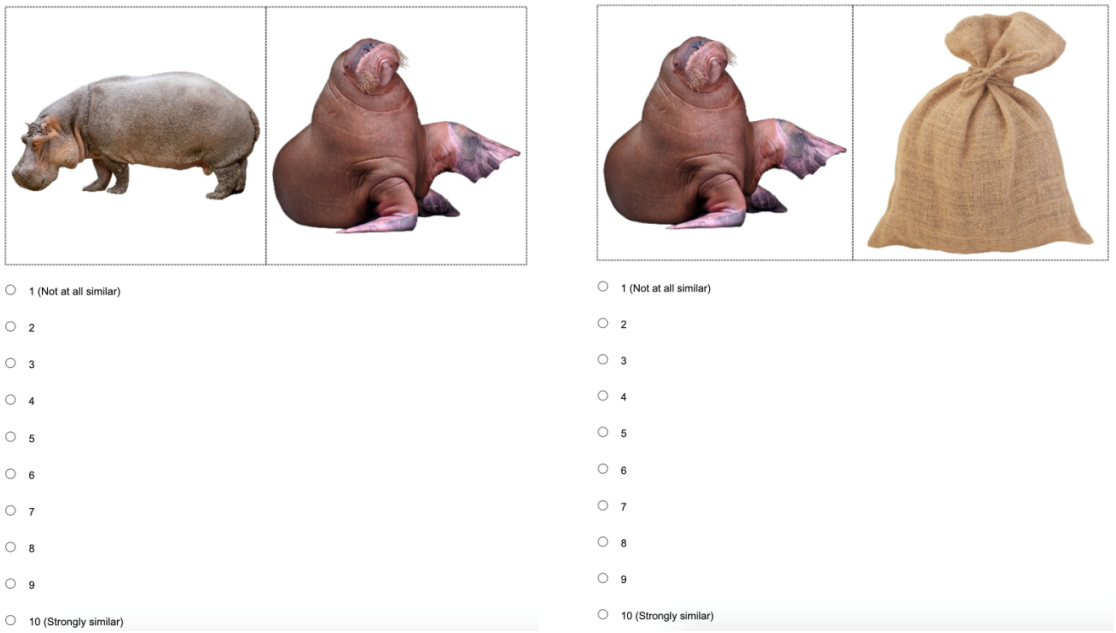


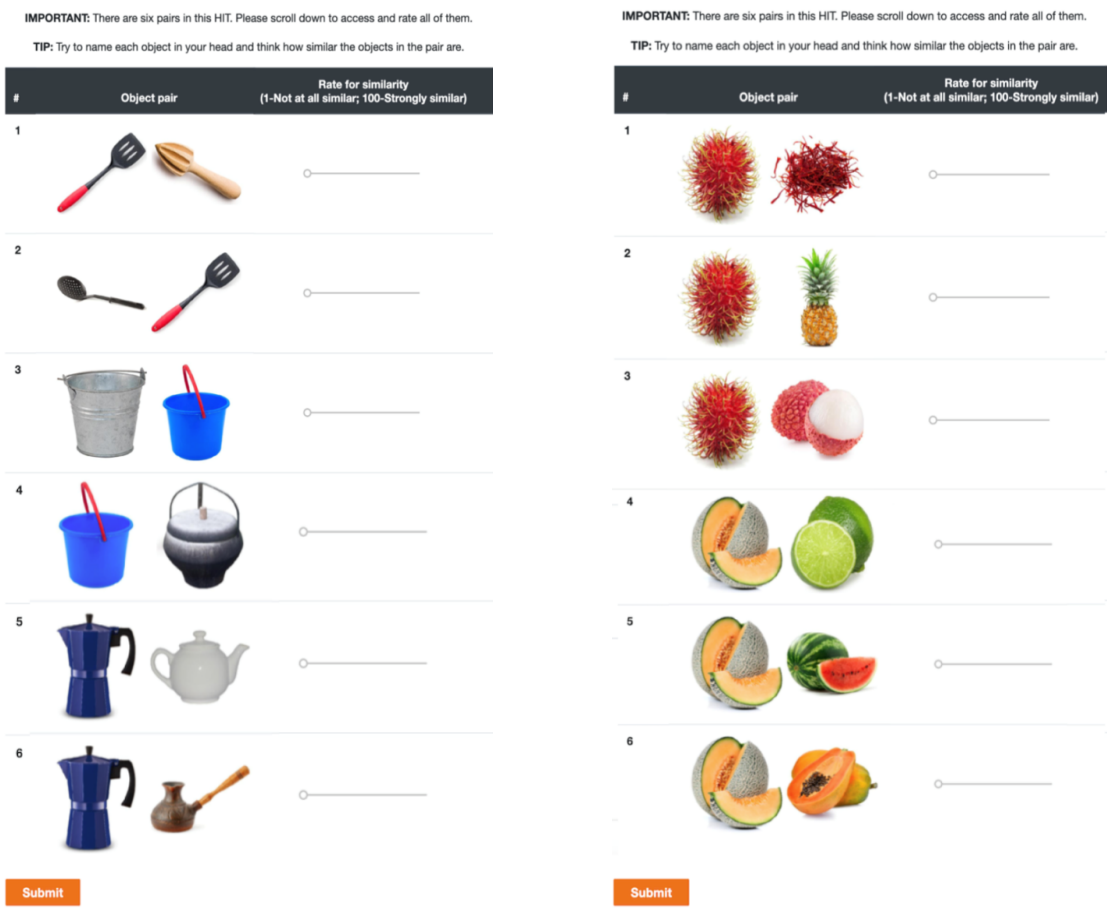
Figure 5.2. *Mandatory qualifying test.*

If the participants fulfilled all the requirements laid out in the qualification test as detailed in this section, they automatically proceeded to the main questionnaire. If not, they could reattempt the qualifying test indefinite number of times till the experiment was live but needed to wait 10 minutes after each failed attempt.

5.3.3.2 Questionnaire group design

Upon qualifying, the participants rated the image pairs in groups of six by moving the slider on the visual analogue scale as shown in Figures 5.3 and 5.4. As discussed earlier, the image pairs were assembled into groups by their semantic categories. A group was said to belong to a semantic category if at least one image in each of the

six pairs belonged to that category. This categorical grouping of image pairs gave context to the participants by facilitating a comparison for their ratings among different image pairs. Two examples of such groups belonging to semantic categories UTENSILS and FRUITS are shown in Figure 5.3. Even within groups, image pairs sharing images were sub-grouped where possible to maximise facilitation of comparison. For example, images of *lime, watermelon* and *papaya* matched with a *cantaloupe* by the computational models were grouped together. This "ranking and paired comparison" style of assessment has been used successfully in past questionnaires such as the Minnesota Importance Questionnaire (Gay et al., 1971) and the O*NET Computerised Work Importance Profiler (McCloy et al., 1999) and has been reported to reduce biases (Sung and Wu, 2018) in subjective participant ratings.



(A) (B)

Figure 5.3. *Example groups rated in the main experiment by participants.*

To rate the similarities, the more granular visual analogue scale (Guyatt et al., 1987; Flynn et al., 2004) was selected over simpler Likert scale (Likert, 1932) for this similarity rating task. Even though Likert scales have been a common choice in the experiments discussed in section 5.2, a major issue with these scales is the ambiguity in selecting the number of response categories (typically three or more). Responses on a Likert scale have also been shown to hide the true latent responses of the participants due to limited selection options (typically 3 to 7), in turn adversely affecting the use of statistical analysis methods and subsequent results (see Allen and Seaman, 2007; Sung and Wu, 2018 for a critique). Greenleaf (1992) reports that based on the response style of a participant, they are likely to choose either extreme or neutral values on a Likert scale, thus leading to biased answers. Additionally, there are difficulties in calibration of the participants responses on the Likert scales. Visual analogue scales help in addressing these shortcomings by providing a much larger range of rating selection, thereby giving the annotators the independence of choosing the rating they believe to be true. The scale used in this experiment was presented as a straight line, ranging from 0 to 100, the two endpoints of which represented extreme values (0: “not at all similar” and 100: “same images”). To continuously remind the participants that this was not a test of visual similarity only, a prompt was given at the top of each group (see Figure 5.3) for the participants to name each object and assess the similarity between object pairs before rating.

Figure 5.4. *Visual Analogue Scale used in the study.*

Upon finishing rating a group, participants were required to submit the results group-wise, as opposed to results of all groups together before proceeding to rating the next group. This was because as per the ethics guidelines as well as in the interest of obtaining reliable results, the participants were allowed to leave the study at any point, and it was not mandatory for them to rate all the groups allotted to them in the questionnaire. It was assumed that the participants' ratings of the groups of image pairs attempted by them were genuine and any unscrupulous ratings could be omitted during the post processing phase (discussed in section 5.3.4). Upon submitting the rating for a group, they automatically advanced to the next group of image pairs. The submitted ratings were readily available and could be accessed using a Python library, Boto3.

5.3.3.3 Questionnaire group split

Following the lead of the experiments discussed in section 5.2, 50 unique ratings per image pair were obtained to allow for more than 30 ratings per image pair post cleaning. A key challenge in obtaining 50 unique ratings for each of the 1,498 image pairs in the stimulus set was to ensure that the annotators did not produce inconsistent ratings due to boredom and fatigue emanating from rating a large number of image pairs. To address this challenge, the 1,498 unique image pairs in the stimulus set were divided into 250 groups of 6 image pairs each. These 250 groups were, in turn, divided into 10 sets of 25 groups each. Each set was made available on the Amazon mechanical Turk platform for 48 continuous hours. Only one set was available at a time on the platform, i.e., there was no overlap with the availability of other sets. This setting ensured that: 1) the participants did not spend any more than the 15-20 minutes on the task needed for rating a set during a 48-hour period and 2) that each image pair was rated only once by a participant, since upon submitting ratings for a group, the participant could not rate that group again. Each set contained a repeat of

one randomly chosen group from the set (also labelled the *consistency group*) to monitor the consistency of participants' responses in the post-processing stage. Participants were paid US $0.10 for each group rated.

5.3.4 Post processing

Upon recording 50 unique ratings for each image pair from the study, the data was post-processed to identify and remove outliers. Similar to previous work by Szumlanski et al. (2013), Silberer and Lapata (2014) and Hill et al. (2015) on similarity ratings, three conditions were imposed for a rating to be included in the final dataset. First, participants who attempted consistency groups needed to have at least a medium agreement between their own ratings on the two instances of the consistency group. This step was necessary to ensure that the participants were at least moderately consistent across the groups they rated in a set (within-set consistency) and did not fluctuate heavily from one group to another. As discussed in detail in the next section, previous studies have measured such agreements using mean pairwise Spearman's rank correlation between ratings on the same image pairs in the two groups. In this scenario, medium agreement was taken as Spearman's $\rho = 0.5$.

Second, if an annotator's ratings for 3 or more image pairs (out of 6) in a group were outliers compared to all other ratings recorded for those image pairs, then their ratings for that group were removed. This ensured the removal of any foul ratings made by the participants at the group-level. Here, outliers were measured using the interquartile range method of outlier detection, which dictates that outliers in a set of data (here, ratings for an image pair) are values 2.5 times interquartile range below the first quartile and 2.5 times interquartile range above the third quartile.

Third, the average pairwise Spearman's rank correlation of ratings by a participant in a group with all other ratings in that group could not be more than one standard deviation below the mean of all such averages (see Hill et al., 2015, p. 678). Any participant whose ratings were below this threshold was removed for that group. This ensured that the participants whose ratings were highly discordant with other annotators in the group were not considered for that group to make the ratings for the group more harmonious.

While incorporating the outlier identification and removal techniques, care was taken that no more than 15 participant ratings should be removed from one group (so that each pair in the final dataset received a minimum of 35 ratings[iv]). The remaining ratings per image pair were averaged to obtain one rating per image pair on a scale of 100. These mean perceived similarity ratings for all the image pairs constitute a new dataset referred to as the ImSim-1498 hereon to aid in further analysis. Examples of perceived image pair similarity ratings from this dataset are shown in Figure 5.5 and the distribution of image pairs by ImSim-1498 similarity ratings is shown in Figure 5.6.

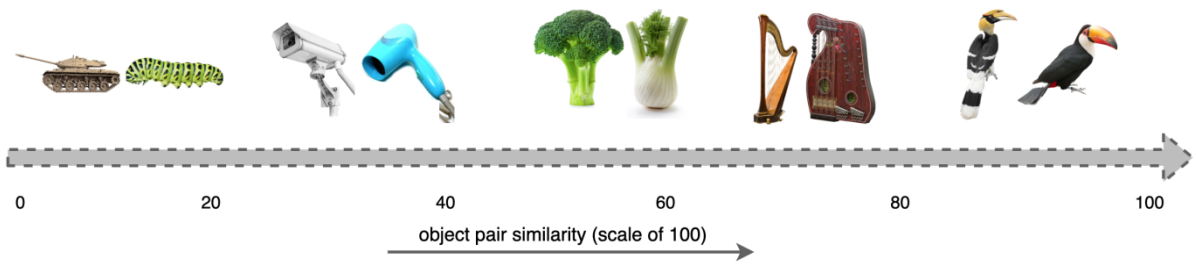


Figure 5.5. *Example images from the new dataset in increasing order of rated similarity.*

[iv] 241/250 groups had 39 or more ratings after post-processing.

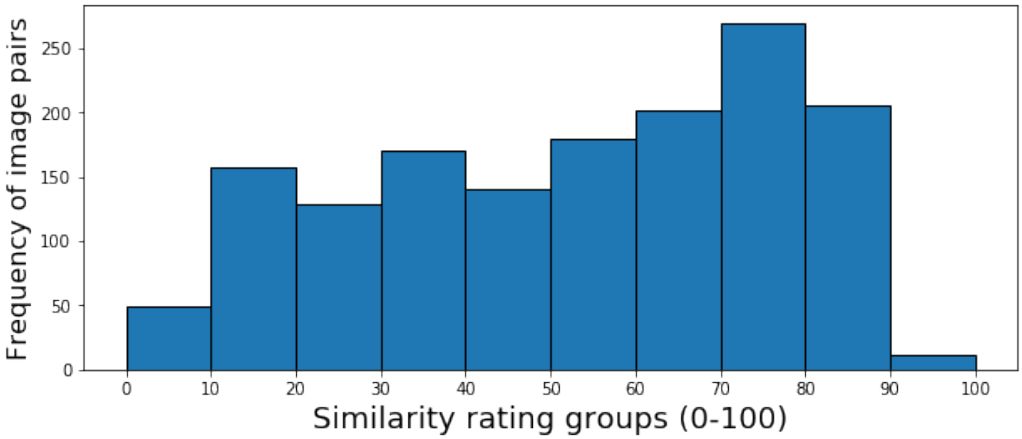


Figure 5.6. *Distribution of image pairs by ImSim-1498 similarity ratings.*

5.4 Statistical Analysis

To address the aims central to this chapter, i.e., to demonstrate the reliability of the ImSim-1498 dataset and to evaluate the computational models against this new dataset, a series of Spearman's rank correlation analyses were conducted. For reference, the interpretation values for Spearman's ρ commonly found in published literature are: 0.2 (small correlation), 0.35 (small-medium correlation), 0.5 (medium correlation), 0.65 (medium-large correlation) and 0.8 (large correlation). As briefly mentioned in the previous section, Spearman's rank correlation has been used in the past studies to measure the agreement between annotators on the ratings they generate for each image pair (see Hill et al., 2015 for more details). The higher this agreement (correlation), the higher the reliability of the dataset. For ImSim-1498, the inter-annotator agreement was calculated as follows. First, the Spearman's rank correlations were calculated for all pairs of unique respondent ratings in each group. The mean of such correlations for each group would give the inter-annotator agreement for that group. Finally, these group-wise agreements were averaged to obtain the inter-annotator agreement on ImSim-1498. Since the same methodology has been used to calculate the reliability of past similarity rating datasets, the comparison of ImSim-1498 with past datasets becomes feasible with this metric. For

comparison of ImSim-1498 reliability, past datasets were carefully chosen to provide legitimate comparisons only. Among the reviewed datasets, WordSim-353, SimLex (concrete nouns) and VisSim were chosen. Among the datasets not used for comparison, MEN dataset was not included because the inter-annotator agreement for the MEN dataset was calculated only on the agreement between the two creators of the dataset and not among all the raters. SemSim was not included because it asked the participants to rate association between objects, which is more abstract than visual or perceived similarity. Finally, the complete SimLex-999 dataset was not included since the agreement on concrete nouns was made available by the authors and would provide a more legitimate comparison with the agreement on the concrete picturable object concepts rated in ImSim-1498.

To evaluate the computational models from the previous chapter against this new dataset, Spearman's rank correlations between the model-generated and ImSim-1498 similarity ratings were also calculated. Similar to the analysis done in the previous chapter, the resulting correlations were used to rank the models against the new dataset to show how well these models capture human understanding of perceived similarity between object image pairs. The resulting rankings of the models on ImSim-1498 were also compared to the rankings obtained on previous datasets to demonstrate how the model estimation of perceived image-pair similarity from ImSim-1498 differed from that of word-pair similarity from the previous datasets.

Finally, to assess the usability of the computationally generated object pair similarity ratings in memory assessment tasks, the assessment technique proposed by Hill et al. (2015) was employed. Per this technique, the Spearman's rank correlations between the model-generated and the ImSim-1498 similarity ratings should be compared to the average of the Spearman's rank correlations between each annotator's ratings and the corresponding ImSim-1498 ratings. A calculation of the latter would give an indication

of how well an average annotator would perform when evaluated against ImSim-1498. The authors suggest that comparing the performance of computational models against this metric would be appropriate (Hill et al., 2015, p. 672) to assess how close a model's similarity ratings are to the systematically acquired average ratings from human participants, thereby providing a quantification of their reliability in downstream analysis. For this analysis, the Spearman's rank correlations between the individual ratings of all object pairs rated by each of the 554 participants and the corresponding ImSim-1498 similarity values for those pairs were calculated. Following this, the average of all such Spearman's rank correlations was taken to obtain the average annotator's Spearman's rank correlation with the ImSim-1498 similarity ratings.

5.5 Results

5.5.1 Inter-annotator agreement

The overall inter-annotator agreement on ImSim-1498 was $\rho = 0.63$ ($p < 0.001$). This medium-large correlation demonstrates that the participants understood the instructions to the task clearly and produced similarity ratings with a reasonable level of consistency. This agreement also compares favourably with the overall agreements of the past datasets VisSim, SimLex (Concrete nouns) and WordSim-353 as shown in Figure 5.7.

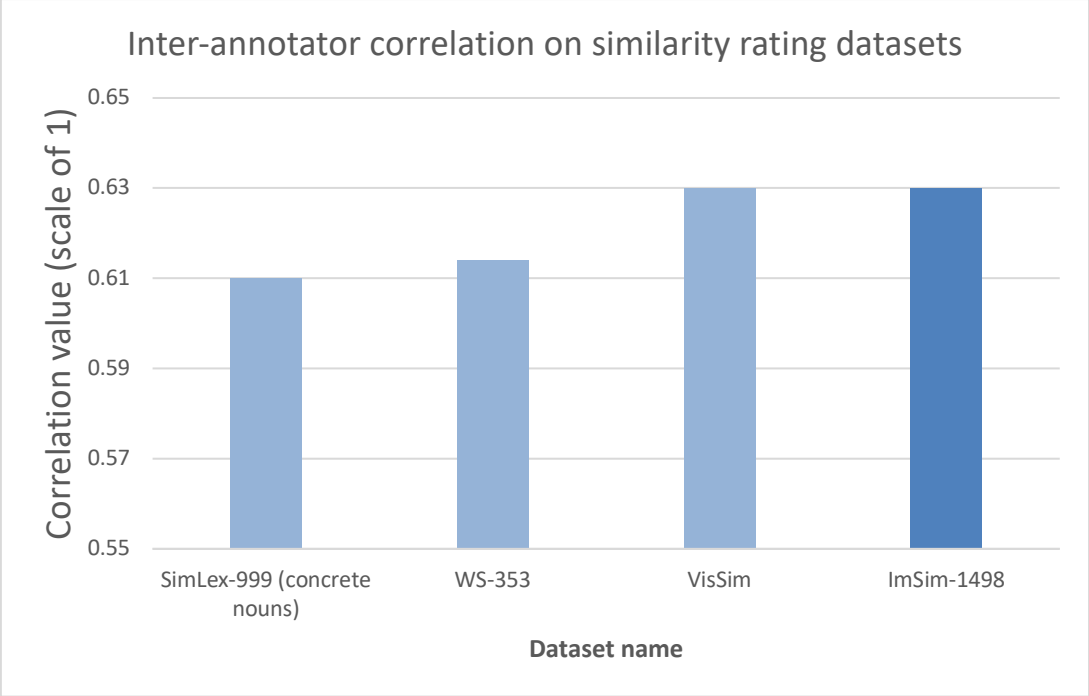


Figure 5.7. *Inter-annotator correlation on similarity rating datasets.*

5.5.2 Evaluation of computational models against ImSim-1498

Spearman's rank correlations between the computational model-generated and ImSim-1498 similarity ratings are reported in Table 5.3, along with the correlations of model-generated similarities with word-based similarity rating datasets from the previous chapter. The visuo-taxonomic model demonstrated the highest correlation (ρ = 0.577), and the linguistic model reported the lowest correlation (ρ = 0.157) with ImSim-1498.

Table 5.3. *Summary of correlation scores between computational models and ImageSim-1498.*

Model	ImSim-1498 (ρ)	SimLex-999 (ρ)	VisSim (ρ)	SemSim (ρ)
Visual	0.42	0.12	0.45	0.45
Linguistic	0.157	0.29	0.46	0.64
Taxonomic	0.467	0.33	0.53	0.68

Visuo-linguistic	0.4	0.36	0.54	0.68
Visuo-taxonomic	0.577	0.36	0.57	0.69
Retrofitted Visuo-linguistic	0.563	0.49	0.62	0.77

A change in the rankings of computational models from performance on word-based datasets to performance on ImSim-1498 is shown in Figure 5.8. As shown, the rankings of computational models remained the same across word-based datasets. On evaluation against ImSim-1498, however, while the visual, taxonomic, and visuo-taxonomic models moved up the rankings, the linguistic, visuo-linguistic, and retrofitted visuo-linguistic models moved down compared to their rankings against word-based datasets. Noticeably, all the models moving down the rankings on evaluation against ImSim-1498 had a linguistic component. To summarise, the take-home message of table 5.3 is that while the models with a linguistic component fared better on the word-based datasets (SimLex-999, VisSim and SemSim), where the participants rated the similarity between concepts without the availability of any images (and only words were shown), these models fared worse on ImSim-1498 dataset, where the images of the concepts were rated for similarity. This demonstrates that the linguistic component of the computational information about a concept from sources such as Word2Vec becomes less important when we are trying to capture the representations of objects based on their similarity rated on the basis of their pictures over mere words.

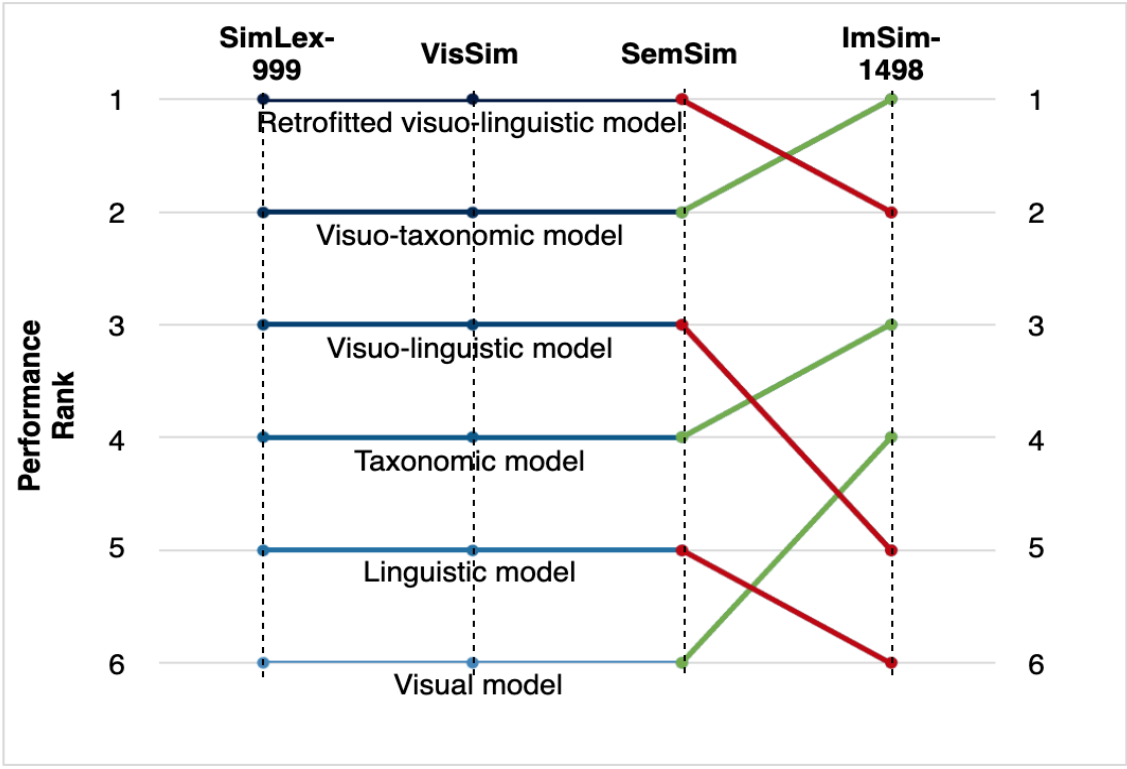


Figure 5.8. *Change in the rankings of computational model performances from word-based datasets to ImSim-1498.*

5.5.3 Evaluation of the usability of computationally generated similarity values

The distribution of the Spearman's rank correlations between each annotator's rating and the ImSim-1498 similarity ratings are shown in Figure 5.9. The distribution shows that a large majority of the annotators had a correlation in the range between 0.7 and 1.0 with ImSim-1498 ratings. As such, the Spearman's rank correlation between average annotator's rating and the ImSim-1498 similarity ratings was calculated as the average of this distribution and was found to be ρ = 0.8144. This correlation value is much higher than the correlation value (ρ = 0.577) obtained for the best performing computational model (the visuo-taxonomic model) on ImSim-1498 in this thesis.

It is worth discussing the outlier correlation values (correlations with ρ = 0.4 and less) in figure 5.9. Despite the post-processing and cleaning of participant ratings before analysis, there could be a few non-adherent participants whose ratings were not removed, leading to a few annotators with correlations of 0.4 or less with the ImSim-

1498 ratings. This can be attributed to the strict post-processing rules detailed in section 5.3.4, whereby ratings by some participants, for example, those who attempted very few groups of image pairs and then quit the experiment for reasons such as boredom, discontinuation of interest in the study, or other miscellaneous reasons could not be removed. It is worth noting that less than 10 such annotators out of the 554 total annotators (< 2%) had outlier correlations with ImSim-1498.

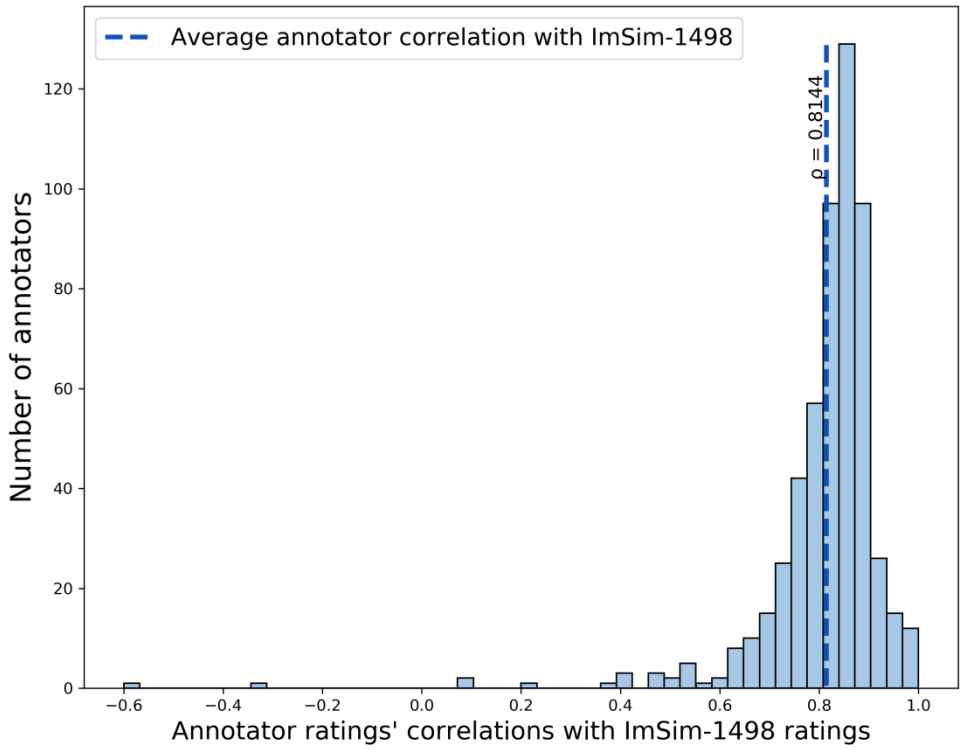


Figure 5.9. *Distribution of Spearman's rank correlations between participant and ImSim-1498 ratings.*

5.6 Discussion

The central aim of this chapter was to obtain highly compliant human-annotated similarity ratings of the computationally matched object pairs from the previous chapter for a methodical assessment of the usability of these pairs in cognitive assessment tasks. To address this objective, a new questionnaire was designed upon reviewing

the best design practices from existing similarity rating questionnaires. The newly designed questionnaire implemented the identified traits in several ways. First, the questionnaire clearly differentiated similarity from association for the participants by providing definitions and examples to explain these concepts (e.g., *bread* and *butter* are related, but not similar, while *bread* and *toast* are similar). The differences between visual, semantic, and overall similarity were also explained to the participants. Second, since the questionnaire provided pictures instead of words, the possibility of wrong sense selection of objects (e.g., *crane* the bird vs. *crane* the machine) was mitigated. Third, the questionnaire included a training, a qualifying test and checkpoint questions to increase the consistency and reliability of similarity ratings. The results from these checks were used to filter and remove unreliable participant ratings. Fourth, the drawbacks of Likert scales were addressed by using the visual analogue scale. Finally, the questionnaire implemented the contextual grouping of object pairs. even beyond grouping, the participants were shown multiple pairs with shared concepts to prompt comparison while rating and ranking the image pairs with the aim of reducing biases and facilitating ease of rating for the participants.

By hosting the designed questionnaire on the Amazon mechanical Turk platform, ratings from a large and demographically diverse sample of 554 participants were obtained. Even after methodical cleaning of these ratings to remove unreliable ratings, it was ensured that no less than 35 ratings per image pair remained. To assess the reliability of the resulting dataset, named ImSim-1498, the inter-annotator agreement was calculated. As expected, this agreement (ρ = 0.63) compared favourably with that of similar datasets and demonstrated that the participants understood what the task expected of them and largely agreed on the ratings.

Upon establishing the reliability of the dataset, the performances of all computational models were calculated on ImSim-1498 and compared with the performances on the

word-based similarity rating datasets from the previous chapter. It was observed that among the computational models used in this thesis, the models with a linguistic component ranked worse on ImSim-1498 than on the word-based datasets from the previous chapter. On the other hand, the rankings of the visual, taxonomic, and visuo-taxonomic models improved on ImSim-1498 compared to their respective rankings on word-based datasets. This change in model rankings could be explained by the nature of ratings that the respective datasets captured. The word-based datasets did not provide pictures and could therefore be capturing relationships between concepts that humans learn from their linguistic environment (see Perfetti, 1998; Barsalou, 1999 and Glenberg and Kaschak, 2002 for support), which is why models with a linguistic component must be performing better on these datasets. On the other hand, the ImSim-1498 dataset asked the participants to rate specific images of concepts and therefore must be capturing the visual component of similarity more than the word-based datasets do. It therefore appeals to the intuition that computational models without a linguistic component would rank lower on ImSim-1498 than those heavy on visual information.

As noted, the visuo-taxonomic model was the best performer on ImSim-1498. That this model outperformed the visuo-linguistic (and even the retrofitted visuo-linguistic) model on ImSim-1498 indicates that when combining semantic and visual information to estimate human knowledge of perceived object image pair similarity, taxonomic input provides a better proxy for semantic information than linguistic input. As iterated in previous chapters, this could be because while linguistic feature vectors are founded on co-occurrence of words in text and therefore capture associative information (e.g., concept *apple* could be linguistically close to concepts *pear, orange,* and *pineapple,* but also to concept *knife*), taxonomic similarities are based on grouping of lexicons (e.g., all mammals grouped together) and hence are more likely to contribute information that assists in grouping perceptually similar objects together. In fact, while

retrofitting the visuo-linguistic model with taxonomic information undid some effect of associativity from the linguistic model, it couldn't remove this effect completely, thereby resulting in a model that performed worse (albeit minutely) than the visuo-taxonomic model on ImSim-1498.

Finally, to address the usability of the model generated similarity ratings of object pairs, the Spearman's rank correlations between the model-generated and the ImSim-1498 similarity ratings were compared to the average of the Spearman's rank correlations between each annotator's ratings and the ImSim-1498 similarity ratings. While past literature has sometimes compared the Spearman's rank correlations between the model-generated and the human similarity ratings with the inter-annotator agreement (Padó, et al., 2007; Reisinger and Mooney, 2010; Silberer and Lapata, 2014), Hill et al. (2015, p. 672) have argued that this is not an appropriate (like-for-like) comparison. The authors argue that "*because the model is compared to the gold-standard average across all annotators* (sic), *we should compare a single annotator to the gold-standard average over all annotators*" to obtain a comparable human metric, where, by "gold standard", the authors mean the cleaned and averaged single ratings per object pair as those in ImSim-1498. In other words, since the model ratings are compared to the ImSim-1498 ratings, a fair comparison of model performance would be against the average annotator performance on ImSim-1498 and not against the inter-annotator agreement. In such a case, it can be easily seen that the performance of even the best performing model (ρ = 0.577) from those used in this thesis is far from the average annotator performance (ρ = 0.8144) on ImSim-1498. While no record of the impact of the different levels of differences between these two metrics was found in literature, the difference here is large enough to assume that even the best performing model employed in this thesis is some way from automatically producing object pair similarity ratings that can be trusted with the same confidence as their manually rated equivalents. Since the computationally produced object pairs were to be used in a

memory assessment task in this thesis to sensitively record the cognitive differences among adults on their performance across the spectrum of object pair similarity, it was concluded that it is more appropriate to use the similarity ratings of such object pairs from ImSim-1498 for this purpose. Before concluding this discussion, it is imperative to mention that some may also argue that computational model performance should only be compared to a Spearman's ρ value of 1.0, since only such a comparison would reflect how well a model performs against an ensemble of humans, which is the equivalent of the "gold standard" of object pair similarity in this thesis. In such a case too, the above argument holds that even the best performing computational model is some way from automatically producing object pair similarity ratings that can be trusted with the same confidence as their manually rated equivalents.

5.7 Conclusion

In this chapter, I presented a new dataset of object pair similarities, the ImSim-1498, that provides favourably agreed upon ratings of the object pairs computationally matched in the previous chapter. The dataset provides similarity ratings for 1,498 object pairs parametrically varying in similarity, thereby adding to the arsenal of the existing concept similarity databases. As demonstrated in the next chapter, this database is useful in assessing the differential effects of object pair similarity on interference in short term memory in adult participants belonging to different demographics, thereby assisting in a deeper understanding of human cognition and behaviour. Such knowledge is critical for early detection of diseases such as AD through early discovery of behavioural and cognitive changes in individuals.

Chapter 6

Assessing the utility of computationally matched object pairs in an objective cognitive task

6.1 Introduction

In the previous chapter, I showed how the human-rated similarity ratings for the object pairs produced by computational models were obtained. In the present chapter, I assess whether these matched pairs inflict confusability when used in an objective cognitive test and whether increasing object pair similarity differentially impacts participants from different demographics, thereby examining the possible use of this metric for sensitively recording cognitive differences in healthy adults. To answer these questions, the confusable object pairs obtained and validated in previous chapters were deployed in a paired associate learning task called Gallery Blast, which was remotely attempted by a large sample of UK-based participants with varying demographics under the GameChanger study.

Gallery Blast is a paired associate learning task like the Mezurio Gallery Game (discussed in Chapter 1). It leverages on the interference theory of forgetting, which states that a memory can be affected at the time of or immediately after learning by other similar interfering memories (see for recent support: Loewenstein et al., 2003, 2004; Wixted, 2005; Dewar et al., 2012; Cybenko, 2011; Curiel et al., 2016). This interference can be of two types: proactive and retroactive. Proactive interference (PI) occurs when intrusion from an old memory prevents successful learning and retrieval of a new memory. Correspondingly, retroactive interference (RI) occurs when intrusion from a new memory prevents successful retrieval of an old memory.

While earlier research (Broadbent, 1958; Murdock, 1964) assumed that memory failure in short-term memory with age or in early AD is solely due to a reduction in the working memory's capacity to hold information, theories such as the inhibition-deficit theory (Hasher and Zacks, 1988) and the competition-response theory (McGeoch, 1942) have shown that memory failure more likely occurs because of failure to inhibit similar memories when retrieving target memory at recall, as suggested by the interference theory of forgetting. Tasks leveraging on these theories (e.g., LASSI-L: Loewenstein and Acevedo, 2005; MITSI-L: Curiel et al., 2016; IPT, IES-R: Woud et al., 2019) have confirmed that testing the failure to inhibit PI and RI from similar stimuli in short-term memory is a more effective indicator of early AD as well as age-related decline in cognition as compared to testing general reduction in short-term memory capacity. The Gallery Blast task central to this chapter leverages on this theory to test susceptibility to PI and RI among its participants.

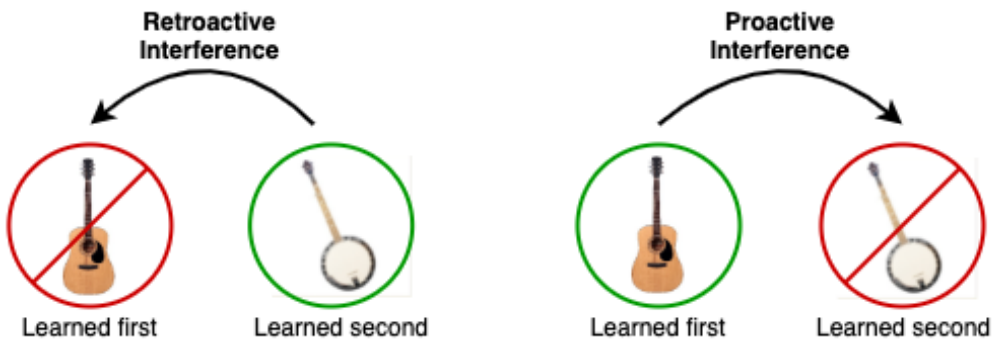


Figure 6.1. *RI: The disruptive effect of new learning on recall of old learning. PI: The disruptive effect of old memory on new learning.*

Traditionally, the tasks measuring susceptibility to PI and RI employ an experimental technique famously called the AB/A'C paradigm (McGeoch and McDonald, 1931; Baddeley and Dale, 1966; Saltz and Hamilton, 1967; Martin 1971a, 1971b; Antony and Bennion, 2020) that works as follows: stimulus A (e.g., *guitar*) is paired with one associate B (e.g., an associated direction) and A', which is similar but distinct to stimulus A (e.g., *banjo*), is paired with a different associate C (a direction different from

B). Participants first learn the A-B association and are then asked to learn the A'-C association. PI occurs when during A'-C learning, the working memory is not able to strongly bind A' with C due to competition from the A-B association learnt earlier. Similarly, RI occurs when the participants are asked to recall the initial A-B association, which is affected by the competing A'-C association at recall. As with most aspects of memory, interference has been shown to be influenced by several factors including age, cognitive capacity, and similarity between competing stimulus pairs (Ebert and Anderson, 2009; Shimamura et al., 1995; Erickson et al., 2014). Regarding stimulus pair similarity, as shown in Chapter 2, it is generally accepted that increasing similarity between old and new stimuli increases interference (similarity between A and A' in AB/A'C paradigm; see also Robinson, 1920; Robinson, 1920, 1927; Skaggs, 1925; McGeoch and McDonald, 1931; McGeoch, 1942; Wixted, 2004; Picker, 2015). However, studies examining differential effects of stimulus pair similarity on RI and PI have found this theory to not 'always' be true (see Cybenko, 2011 for a review). While there seems to be more agreement that this theory is likely true for susceptibility to RI (see Skaggs, 1925; Lund, 1926; Robinson, 1927; Cheng, 1929; Osgood, 1949; French, 1999; Anderson et al., 1994; 2003; Wixted, 2004; Cybenko, 2011; Darby and Sloutsky, 2013; Picker, 2015; Antony and Benion, 2020), literature is divided on the relationship between similarity and PI (see Bennet, 1975; Samrani et al., 2017 for a positive relationship between PI and similarity; and Chanales et al., 2019; Antony and Benion, 2020 for a negative relationship). This is discussed in detail in section 6.5. Notwithstanding the relationship between interference and the similarity between competing memories, studies have generally found RI effects to be stronger than PI effects among their subjects (Schmeidler, 1939; McGeoch and Underwood, 1943; Melton and Von Lacrum, 1941; Underwood, 1945; Curiel et al., 2013; Crocco et al., 2014). Cybenko (2011, p.15) summarises this phenomenon as follows: "*material presented to a participant after the target information is presented is more detrimental to their memory accuracy than material presented to a participant prior to the target*

information". Apart from answering the questions central to this chapter, the findings from the Gallery Blast task analysis also have the potential to contribute to these ongoing debates on the differential strengths of PI and RI effects, as well as the differential effects of stimulus pair similarity on PI and RI in short term memory.

However, to address the central aims underlined earlier in this section, i.e., assessing the utility of the computationally generated object pairs in the Gallery Blast task and examining whether the object pair similarity metric can be used for sensitively recording individual cognitive differences, I divided the analysis into four parts. First, I examined the validity of Mezurio Gallery Blast as a reliable test of cognition. This was done by finding the relationships between demographic characteristics, not limited to age, education level, and family history of AD and participant performance on the task from a large sample of UK-based participants with varying demographics and no diagnosis of cognitive impairment. The expectation was that if the task is valid, these relationships would align with those widely accepted in literature (e.g., more susceptibility to interference with age, or less susceptibility with higher maximum education attained; detailed in section 6.5). Second, to assess the utility of computationally generated confusable object pairs in the Gallery Blast task, I examined whether these object pairs inflict interference in short-term memory when deployed in the task. The expectation was that the participants would experience both RI and PI, but RI effects could be stronger than the PI effects, as discussed earlier. Third, the specific effects of object pair similarity on intrusion errors due to both RI and PI were obtained. It was predicted that increasing the object pair similarity would increase intrusion errors due to RI, but no prediction was made for the PI condition given limited agreement on this relationship. Finally, the differential effects of the object pair similarity metric on healthy adults belonging to different age groups were found to examine the potential use of this metric for sensitively recording individual

cognitive differences in the Gallery Blast task, thereby providing support for the potential validation of this metric against traditional biomarkers of AD.

6.2 Materials and Methods

6.2.1 Participants

The participants for the analysis described in this chapter were selected from the second round of the GameChanger study (Lancaster and Hinds, 2019). GameChanger is a longitudinal e-cohort open to the general UK public. A major shortcoming of typical lab-based studies is the recruitment of small samples of participants that lack diversity, thereby limiting the generalisability of the findings from such studies. For instance, Güsten et al. (2021) argue that due to the inability to recruit participants from varied demographics, a large majority of studies examining cognitive decline need to categorically divide participants into young and old despite the underlying decline being shown to be a fundamentally continuous process (Nyberg et al., 2012). By making objective cognitive tasks from the Mezurio platform available remotely, GameChanger addresses this major shortcoming of lab-based recruitment. The aim of GameChanger is to establish normative performance on the tasks in Mezurio in a wide demographic with the goal of using the obtained findings as baseline when considering high-risk cohorts from other studies. The study aims to capture longitudinal participant performance on Mezurio tasks over three successive rounds, with two rounds already completed. To participate, volunteers must be aged 18 years or older, not have a dementia diagnosis and have daily access to an Apple or Android smartphone. Recruitment of participants to the GameChanger study has primarily taken place via Alzheimer's Society media channels and the Join Dementia Research database.

After excluding participants based on the criteria reported later in this section (section 6.2.6), a sample of 1,455 adults (aged 19-95 years as on March 1, 2020; 77% female)

was analysed. Their descriptive data is shown in Table 6.1. Among the 1,455 adult volunteers, 52.16% reported a family history of AD, which is significantly higher than the proportion reported in a prospective population study (18.6%) (Qian et al., 2017). The study protocol was ethically approved (University of Oxford Medical Sciences Inter-Divisional Research Ethics Committee: R58202/RE001). E-consent via the GameChanger website (https://joingamechanger.org) was required upon study sign-up. All procedures were in accordance with the Helsinki declaration.

***Table 6.1**. Demographic characteristics of participants from GameChanger round 2 selected for analysis.*

Demographics	
Overall *n*	1455
Age (years) (*M* ± SD)	58.89 ± 11.16
Number of participants by age group	
18-30	34
30-40	61
40-50	153
50-60	438
60-70	539
70-80	211
80+	19
Biological sex (% Female)	79.66
Immediate family history of dementia (%)	52.16

Highest education level (%)	
Primary school	0.0
Secondary school	12.23
College	13.95
Trade or Technical	10.45
Undergraduate	38.56
Postgraduate	24.26

6.2.2 Procedure

Following registration and e-consent on the study website, participants received an automated email, which included a weblink to download the Mezurio application as well as a unique ID to sign into the app. After signing into the app, each participant was asked to provide demographic information including but not limited to age, gender, education level and family history of dementia. Following this, the participants were prompted to complete an array of tasks within the Mezurio app for 31 days (approximately 5 minutes each day). Apart from the Gallery Blast task central to this thesis, such tasks included measures of episodic memory (Gallery Game discussed in Chapter 1), visual processing speed, executive function and language, alongside self-report scales of mood, sleep, and subjective cognition, each purpose developed for remote, smartphone administration (see Lancaster et al., 2019 for details).

In the last 7 days (days 25-31) of the 31-day GameChanger study, the participants were required to attempt the Gallery Blast task for approximately 5 minutes each day. The app encouraged participants to play at the same time each day by sending a phone-based notification, with a reminder notification sent 15 minutes later if the task

was not initiated. Participants had 16 hours to complete each activity after receiving the prompt; tasks not completed within this time window 'expired'. Since the Gallery Blast task was inspired by Gallery Game and was being introduced for the first time to the general public in GameChanger round 2 (2020), it was only made available to those participants who had attempted the Gallery Game task before. This ensured that the participants had had an opportunity to practice tasks in common with the Gallery Game and had enough experience of forming their own strategies to assist learning before attempting the Gallery Blast task, thereby minimising individual differences at encoding for a non-confounded interrogation of individual differences in susceptibility to interference.

6.2.3 Stimuli

The human-rated, machine-generated confusable object pairs from the previous chapters were used as input to the Gallery Blast task. Two independent researchers involved in the design of the experiment individually selected and agreed upon the similarity rating threshold of 42/100 from the ImSim-1498 dataset, above which the object pairs shared semantic categories and hence were highly similar. 126 of these object pairs were chosen for use in Gallery Blast – 42 image pairs for each new cohort joining in the three intended years of the GameChanger study. These pairs were carefully chosen upon giving due consideration to their semantic categories (discussed further in section 6.2.4) and their spread over the object similarity scale to obtain generalisable effects of object pair similarity on the outcome variables. For the analysis presented in this chapter, 84 of the 126 such image pairs are considered, since this chapter only assesses data from year 1 and year 2 participants in GameChanger.

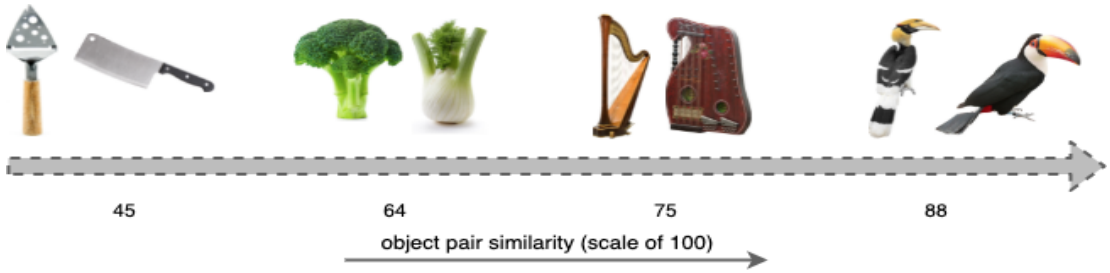


Figure 6.2. *Examples of stimuli used in the study.*

6.2.4 Mezurio Gallery Blast

Gallery Blast is a freely available cognitive task, deployed within the Mezurio smartphone app platform (https://mezur.io), designed to capture PI and RI effects from competing memories of perceptually similar objects in the short-term memory. To capture these effects, the task has been adapted from the AB/A'C design discussed in section 6.1. The task has a repeated-measures design, with the participants asked to complete two learning blocks, each associated with multiple recall tests. In a single sitting that typically lasts 5-7 minutes, participants learn 'swipe' directions associated with 6 object images, followed by the directions associated with their confusable counterparts (6 matched images) in separate paired associate learning blocks. Upon learning these associations for objects in both the blocks, participant performance on the recall of the learned object-direction associations demonstrates their susceptibility to PI and RI. All the tasks and tests were attempted by the participants on different sets of confusable image pairs each day.

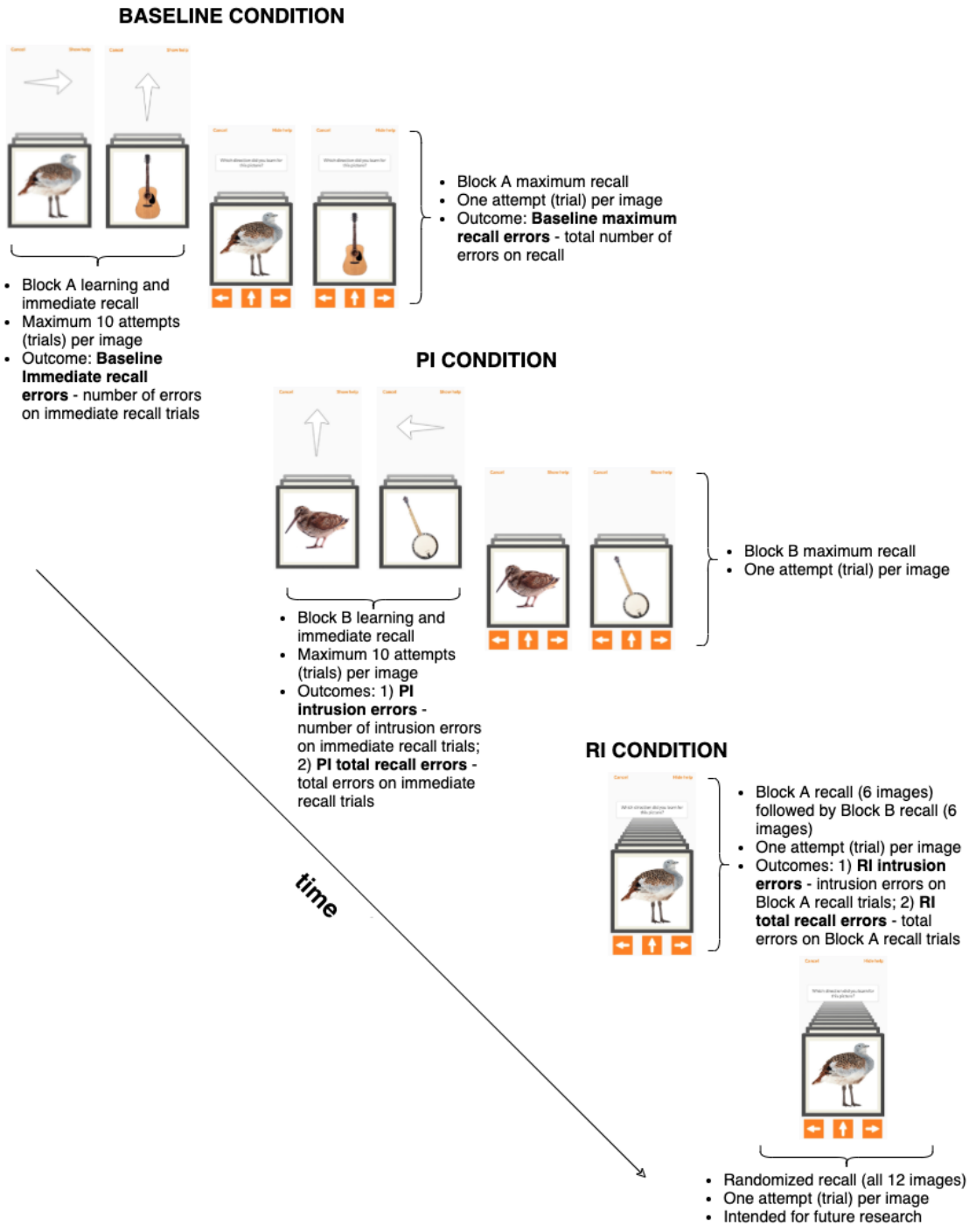


Figure 6.3. *Gallery Blast task workflow.*

6.2.4.1 Task description and outcome variables

The Gallery Blast task started with a learning block of 6 mutually unrelated object images, hereafter referred to as 'block A'. The block was administered as two successive sub-blocks of 3 living and 3 non-living object images, as shown in Figure

6.3. All the living and non-living objects in the block belonged to different semantic categories (e.g., BIRD, MAMMAL, and FRUIT in living and VEHICLE, TOOL, and FURNITURE in non-living). An example of a single learning trial is shown in Figure 6.4A. Each object image was associated with an arrow pointing left, right, or up. In the first iteration (composed of a 'gallery' of 3 object images), participants were shown the object as well as the associated direction and were required to 'swipe' the image in the cued direction to learn the object-direction association. The directional cue was absent on subsequent presentations of this object to test immediate recall for object-direction pairings, with an incorrect response triggering a second object-direction learning trial with the directional cue. Within each subsequent iteration, the images were presented in a random order to avoid confounds of stimulus order. The iterations continued until the individual reached a 'stopping criterion' that 6 out of the last 8 presentations of a given stimulus are responded to correctly, with a maximum of 10 attempts allowed per image. This repeated presentation of object-direction pairings until the criterion for successful learning was reached was aimed at minimising individual differences at encoding for a non-confounded interrogation of susceptibility to interference. The total number of errors made by each participant on the immediate recall trials in block A was recorded as **baseline immediate recall errors**. This variable has been shown to represent the pre-interference global learning deficit of the participant. Past research asserts that controlling for this variable while measuring PI effects on participants is needed to avoid confounds such as effects due to non-AD related impairments and differences in learning capacity.

Upon passing the stopping criterion in block A, all the 6 learned images were presented one at a time (once each) to the participant to test the retrieval of their learned directions. A single recall trial is shown in Figure 6.4C. Participant performance on this recall test provides a measure of their maximum recall capacity of block A images (Curiel et al., 2016). The total number of errors made by each

participant on this task was recorded as **baseline maximum recall errors**. Lancaster and colleagues (2020) assert that in Gallery Game too, participants demonstrate ceiling recall performance immediately after successfully concluding the learning block. Controlling for this variable helps assess true RI effects when block A object-direction pairs are recalled after being subject to interference as described next (Curiel et al., 2016; Crocco et al., 2014; Loewenstein et al., 2003, 2004).

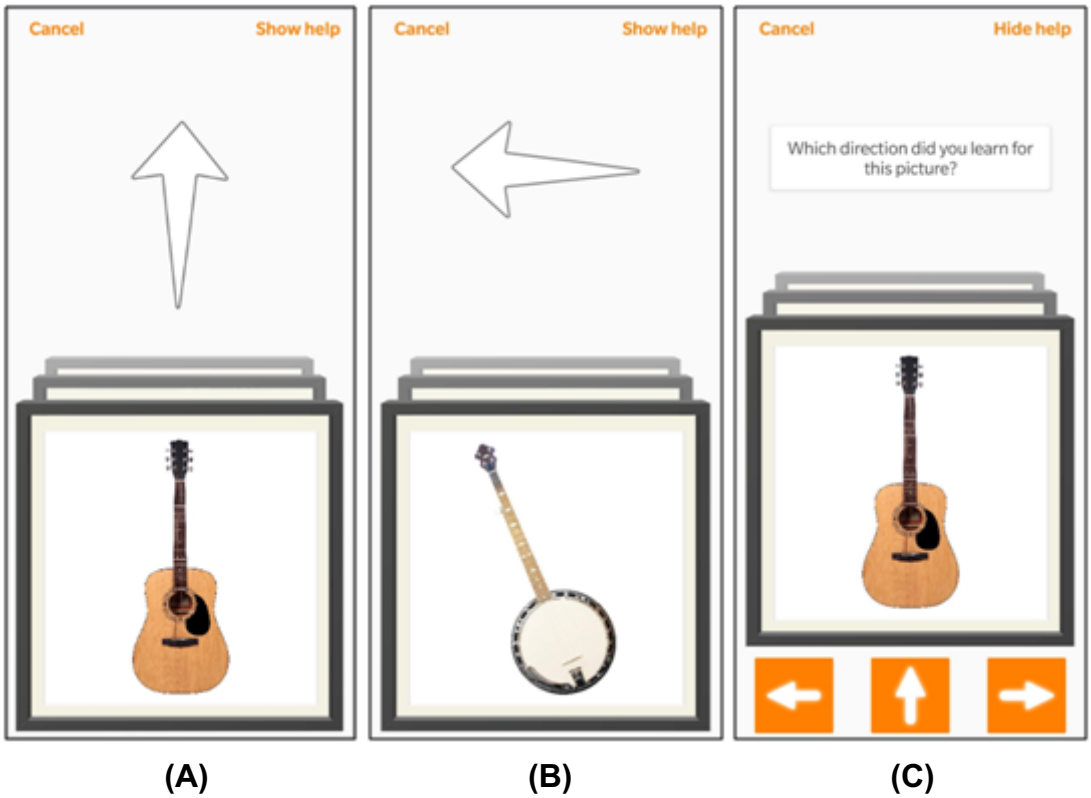


Figure 6.4. *Representation of (A) a single learning trial in block A, (B) a single learning trial in block B and (C) a single recall trial.*

Following this, to assess susceptibility to PI, the participants were introduced to a second learning block (block B) that consisted of 6 other object images that were perceptually similar to those presented in block A (e.g., *guitar-banjo*), again divided into 2 sub-blocks of 3 living and 3 non-living images each. The selection of these stimuli has been discussed in section 6.2.3. It was ensured that the swipe directions associated with the images in block B were different from the directions associated

with their matched counterparts in block A. Block B was administered with the same procedure as block A, i.e., the participants were expected to learn the presented object-direction associations for the new set of images, which were tested for immediate recall in the subsequent presentations in the absence of the directional cue. To analyse the effects of PI, i.e., interference from block A object-direction association memories on the encoding of the new associations in block B, two types of immediate recall errors made per participant per stimulus in block B were recorded: intrusion and total. Intrusion errors are commonly recorded in paradigms investigating PI and RI effects (Curiel et al., 2013; Crocco et al., 2014; Loewenstein et al., 2018) and have been defined in past literature as the class of errors in which the participant retrieves information about a previously encountered similar memory instead of the target memory (Curiel et al., 2013). With regards to Gallery Blast analysis, an intrusion error translates to the incorrect retrieval of the direction associated with the matched stimulus during recall instead of that associated with the target stimulus. Consequently, intrusion errors due to PI in Gallery Blast were calculated as the number of immediate recall trials in block B in which the direction for the stimulus in question was incorrectly marked as that of its corresponding similar stimulus from block A. For instance, from Figure 6.4, upon being asked to recall the direction associated with *banjo* from block B, if the participant wrongly retrieved the direction of *guitar* from block A, then this was marked as an intrusion error. Hereafter, these errors will be denoted as **PI intrusion errors**. Additionally, the total number of errors (intrusion and non-intrusion) made per participant per stimulus on immediate recall in block B were recorded as **total PI recall errors**.

Upon passing the stopping criterion in block B, a single recall block was presented to test the maximum recall capacity of the participants for block B, as shown in Figure 6.3.

Having encoded block B object-direction pairings to their maximum capacity, the participants were asked to recall the directions of the images from the original block A to assess their susceptibility to RI. Each of the 6 images learned in block A was presented only once for recall. To analyse the effects of RI, both intrusion and total recall errors were recorded per participant per stimulus as binary "True" or "False", i.e., respectively error made or not made by the participant for the presented stimulus. Specifically, **RI intrusion errors** were recorded as "True" if the direction for the target stimulus was incorrectly marked as that of its matched similar stimulus from block B and "False" otherwise. **Total RI recall errors** per participant per stimulus were recorded as "True" if any class of error (intrusion and non-intrusion) occurred and "False" otherwise.

In the same block, after testing retrieval on block A, images from block B were presented again to test for recall. However, since effects beyond those from PI and RI would start affecting participant performance here, this sub-block was not analysed for the purpose of this thesis. In a similar vein, there was a final block following this task in which the participants were asked to retrieve object directions for all the 12 images learned in blocks A and B combined, except in this block, the images were presented in a random order. The performance of the participants on this block, too, was not assessed for the purpose of this thesis since the effects experienced by participants in this block too would be beyond the effects of interference recorded in existing literature. In future studies, when biomarker-based evidence of participant's cognitive state is available, it will be interesting to see how recall errors from these tests, in addition to those analysed in this chapter, correlate with such markers.

6.2.5 Variables of interest

The demographic and stimulus-level variables whose effects have been studied and controlled for in the analysis undertaken in this chapter are summarised alongside the outcome variables in Table 6.2. They are described in detail in this section.

***Table 6.2**. Variables of interest for Gallery Blast analysis.*

Variable name	Definition	Type	(M±SD) Range
Outcome variables			
Baseline immediate recall errors *	Pre-interference memory/attention deficit on immediate recall	Numerical	5.83 ± 5.60 [0-49]
Baseline maximum recall errors *	Pre-interference memory/attention deficit on maximum recall	Numerical	0.68 ± 1.36 [0-15]
Total PI recall errors	Total errors made in the PI condition	Numerical	11.65 ± 9.49 [0-64]
PI intrusion errors	Intrusion errors made in the PI condition	Numerical	8.45 ± 7.02 [0-57]
PI intrusion error rate	Ratio of all trials in the PI condition in which intrusion errors were made	Ratio	0.03 ± 0.022 [0-0.15]
Total RI recall errors	Total errors made in the RI condition	Numerical	5.81 ± 4.79 [0-26]
RI intrusion errors	Intrusion errors made in the RI condition	Numerical	4.82 ± 3.99 [0-21]
Demographic-level variables			
Age	Participant's age	Numerical	19-95^
Biological sex	Participant's biological sex	Categorical	[Male/Female/Other]^
Maximum education level	Participant's maximum formal education level	Categorical	[0/1/2/3/4/5]^
Family history of dementia	Whether participant's immediate family member has received a diagnosis of dementia	Categorical	[Yes/No/Not sure]^
SCD score	Index of self-report of cognitive decline (the higher, the more subjective decline)	Numerical	12.77 ± 5.58 [1-50.625]

Stimulus-level variables			
Object pair similarity (scale of 100)	Human-rated similarity between target object and matched counterpart	Continuous	42-89
Congruency	Congruency between object orientation and associated direction	Categorical	[Same, Opposite, Perpendicular, None]

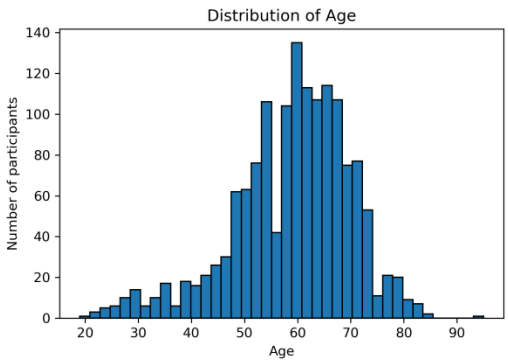
^ see Table 6.1 for details
*-marked outcome variables also considered as demographic-level variables

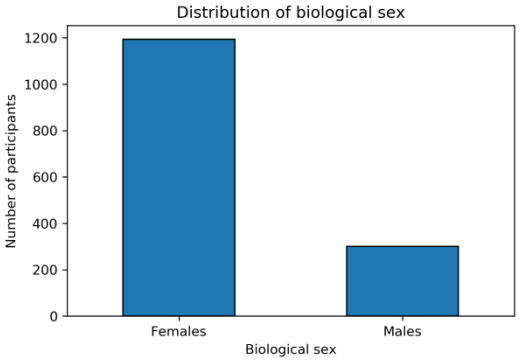
6.2.5.1 Demographic-level

The demographic-level variables recorded from the GameChanger study were age, biological sex, maximum education level, family history of dementia and subjective cognitive decline (SCD). Of these, all but SCD were recorded once at the beginning of the study. As described in section 6.2.2, however, along with other cognitive tasks that constitute Mezurio, SCD was recorded over several days of self-reports on carefully prepared questionnaires consisting of three separate subscales: 1) General Subjective Cognition, 2) Cognitive Slips – 6-months; 3) Cognitive Slips – 24 hours. While the first subscale, designed for infrequent administration (e.g., annual, or once every two years), captured the individual's broader self-perceived cognition, including perception of decline within the past two-years, age-matched comparison, and source of personal worry, the other two captured more frequent cognitive slips hypothesised to increase the validity of self-report by asking participants to reflect on their cognitive function over a narrow timeframe. All questions were multiple choice, with potential answers presented as 'radio-buttons' to ensure participants submit only one response per question. Each response option presents a text-based response (e.g., 'my thinking abilities are a lot worse', 'A little', or 'Yes') alongside a numerical value (e.g., 0 – 4). Prior to producing a single numeric score for each subscale, the numeric values associated with each multiple-choice set were recoded.

For subscales 1 and 2, consistent with the ranked ordinal nature of multiple-choice data, the median score on each item was used as the subscale metric. For the more frequent subscale 3, the performance was summarised as each participant's mean daily total. The addition of these 3 scores provides a Mezurio SCD score, intended to comprehensively assess poor subjective cognition. The internal validity of the scale has been established and future work will include the validation of the scale with established biomarker of very early AD (Lancaster et al., manuscript in preparation). The final SCD score was used as a demographic-level predictor in the analysis of Gallery Blast outcomes to account for self-report of subjective cognitive state for each participant. The higher this score, the worse the subjective cognition.

Apart from the described demographic-level variables, baseline recall errors in both the immediate and maximum recall conditions were controlled for as participant-specific characteristics. Baseline participant performance has been discussed as an important control variable in literature and is needed to account for participant's global memory performance (Crocco et al., 2014; Loewenstein et al., 2018). Crocco et al. (2014) assert that this procedure allows for the participants to serve as their own control and helps account for pre-interference differences in global memory/attention deficit among them.





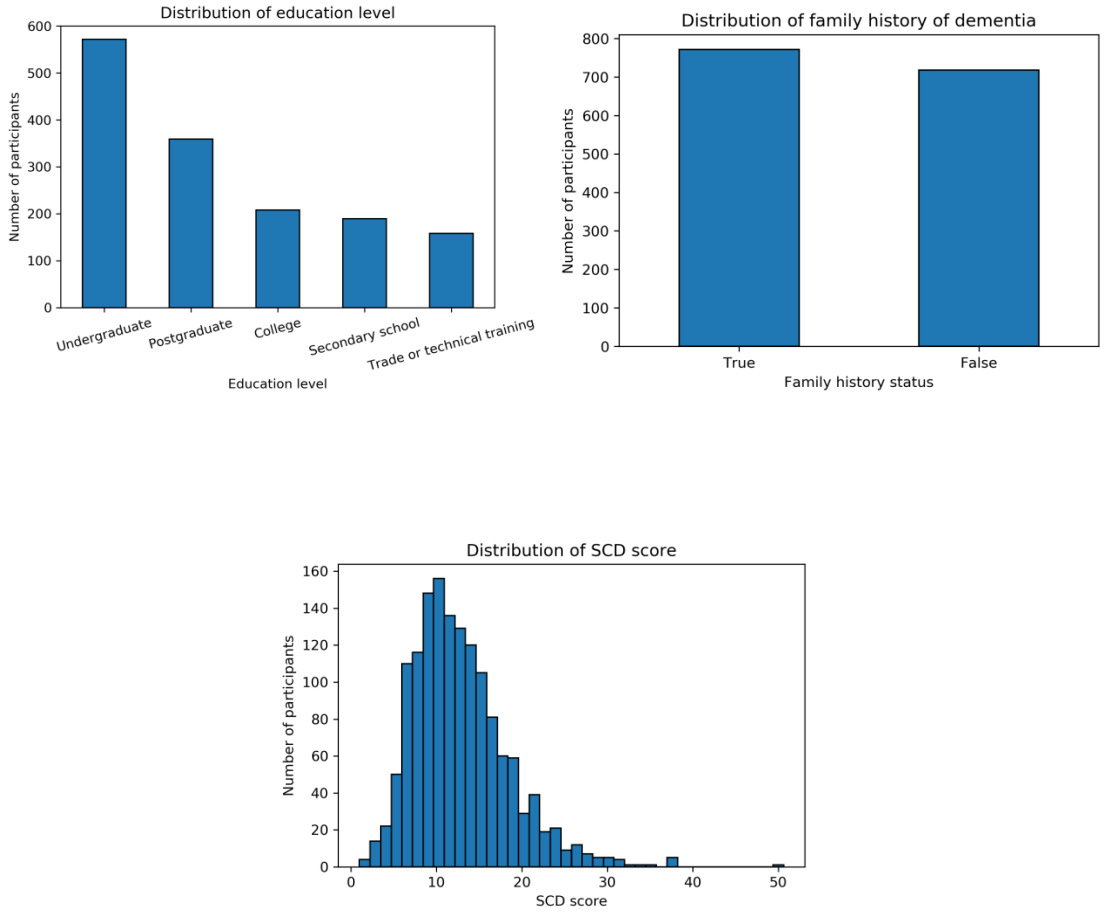


Figure 6.5. *Distributions of the demographic-level variables age, biological sex, education level, family history, and the SCD score. The distribution of age is somewhat normal in that a large distribution of the participants lie in age groups 45 to 75. For biological sex, the distribution is heavily skewed towards females. The education level indicates that the largest proportion of our participants had an undergraduate degree, followed by a postgraduate degree, followed by a college degree, followed by a secondary schooling, followed by trade or technical training as their highest level of education. For family history, our participants were equally balanced between those with and without a family history of dementia. Finally, the subjective cognitive decline scores are almost normal, peaking at 10, with less than 10 participants with a score of 30 or above.*

6.2.5.2 Stimulus-level

The stimulus-level variables recorded for inclusion in the analysis were object pair similarity and congruency. Object pair similarity represents the human-rated similarity between the target image and its computational matched counterpart. It is a continuous variable ranging from 0 to 100, 0 representing completely dissimilar pair of objects and 100 representing extremely similar pair of objects. For the purpose of this chapter, however, the range is limited to [42-89]. The other stimulus-level variable,

congruency, is a categorical variable that represents the 'congruency' of each image's orientation with its associated swipe direction. It has four discrete values: same, opposite, perpendicular and none. The congruency is *same* when the orientation of the object in the target image is the same as the direction associated with the image; it is *opposite* when the orientation of the target image is opposite to the associated direction; *perpendicular* when the associated direction is perpendicular to object orientation and *none* when there is no clear relationship between the object direction and orientation. Controlling for the congruency of the stimulus image is important since it influences the way participants strategize learning the object-direction associations. For instance, in Figure 6.4, since the guitar's head aligns with the associated direction (congruency: *same*), it is easier for the participants to form a strategy to remember this direction than in a stimulus image in which this relationship is *perpendicular*. For the purpose of the analysis undertaken in this chapter, the reference level for congruency was set to *none*.

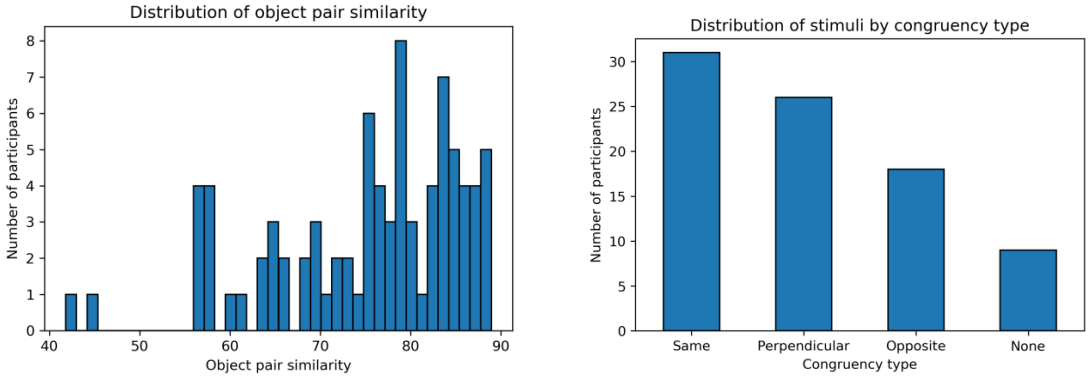


Figure 6.6. *Distributions of the stimulus-level variables object pair similarity and congruency type. The object pair similarity ranges from 42 to 89. The highest proportion of our stimuli had same orientation, followed by perpendicular, followed by opposite, followed by no orientation.*

6.2.6 Participant exclusion

Recent work has shown that remote assessments provide easy access to high quality data. However, following the lead of previous literature (Downs et al., 2010; Thomas

and Clifford, 2017; Güsten et al., 2021), rigorous inclusion/exclusion criteria were applied after data collection to obtain true estimates in this first, exploratory use of Gallery Blast. Gallery Blast was attempted by 2,399 participants in GameChanger round 2. For the analysis presented in this chapter, data from a subset (n = 1,455) of these participants, who attempted the task for five or more of the stipulated seven days, were selected. To select this subset of participants, first, the participants who provided contradicting demographic information (n=227) were excluded. Following this, those who did not provide any demographic information (n=33) were excluded. Following this, participants who attempted the task for less than five of the seven stipulated days of assessment were excluded from the present analysis. The threshold of five days was decided in line with the overall participant adherence in the GameChanger study, which was found to be ~80%. Here, the adherence was calculated as the average number of days the participants attempted the GameChanger tasks as a percentage of the actual study length. Since the present work provides first analysis of Gallery Blast data and is being used to examine the preliminary validation of the task, this cut-off provides truer guarantees of the effects sizes of the variables of interest on the outcome measures than if the data from the least adhering participants was also included in the analysis. Following this, technical glitches in data collection prompted the rejection of another 151 participants. Finally, 39 participants with a self-reported neurological illness, presently or within the past 5 years, were also not included in the present analysis to limit bias in outcomes as a result of non-dementia related impairment (Jessen et al., 2014). It is worth noting here, that exclusion of these participants from the preliminary analysis was agreed upon by our lab group since including data from such participants could pollute the effect sizes that our lab group was interested in obtaining – that for healthy participants who had completed the task at least for a set number of days (as described above), had no neurological illness, and had provided the right demographic information.

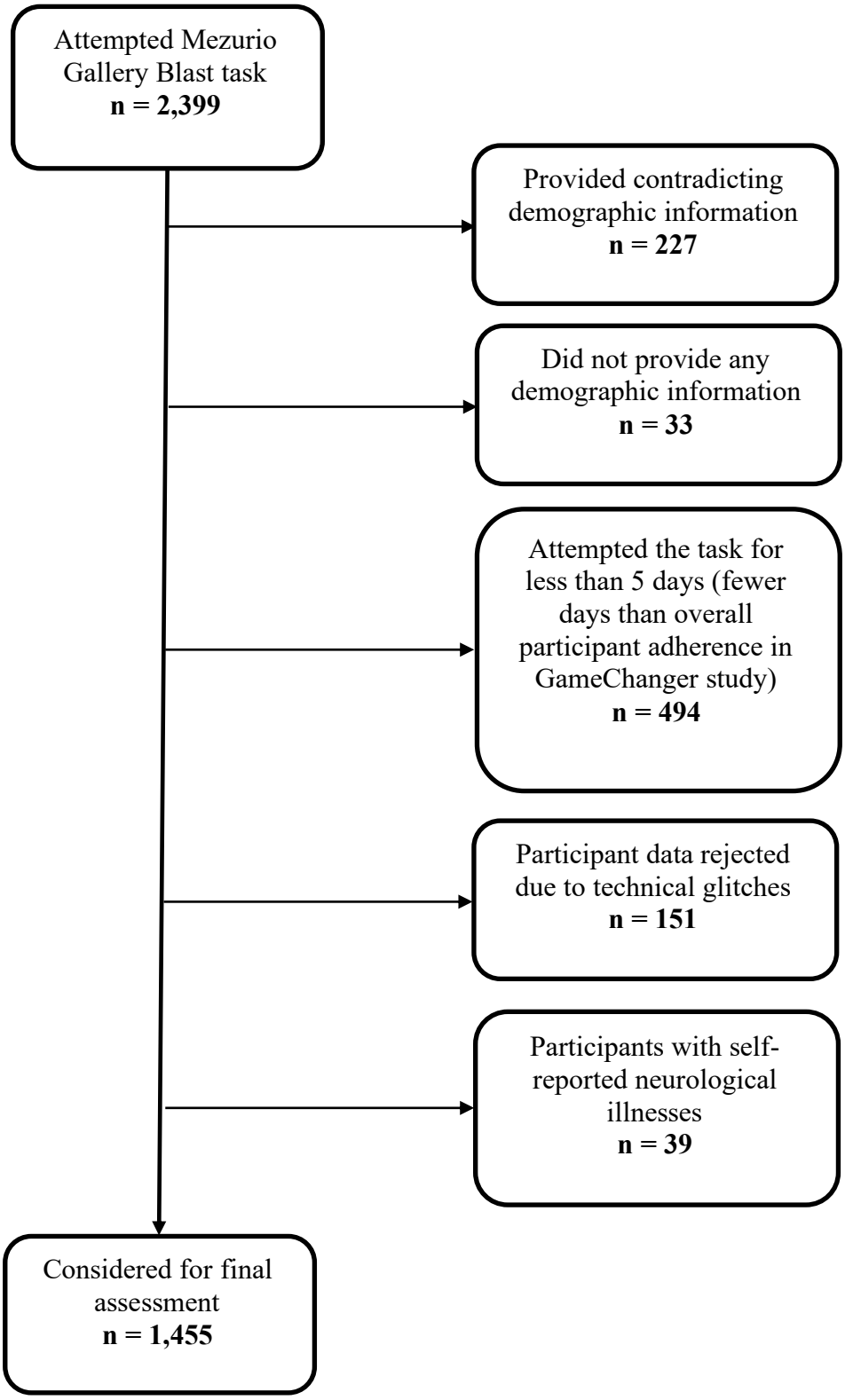


Figure 6.7. *Path diagram showing the number of participants who attempted the Gallery Blast task, who were removed, and who were considered in the present analysis.*

For the purpose of removing participants based on outlier performances, it was decided in advance that only the participants with extreme values on outcome

variables (i.e., those deliberately non-compliant) would be rejected. This was because participant performances in the present study could vary due to a number of demographic factors and hence participants could not be excluded based on simple outlier detection methods. Instead, outliers were manually inspected for foul-play. Participant performances on four key outcome variables (baseline immediate recall errors, baseline maximum recall errors, total PI recall errors and total RI recall errors) are shown in Figure 6.5. Since the distribution is non-normal, the interquartile range (IQR = Q3-Q1) method was used to detect outlier performances (outliers lie beyond Q3 + 2.5xIQR). These are marked with vertical lines in Figure 6.5. Even after adjusting for the number of days that the participants attempted the task, there were no grounds to reject any remaining participants based on outlier performance.

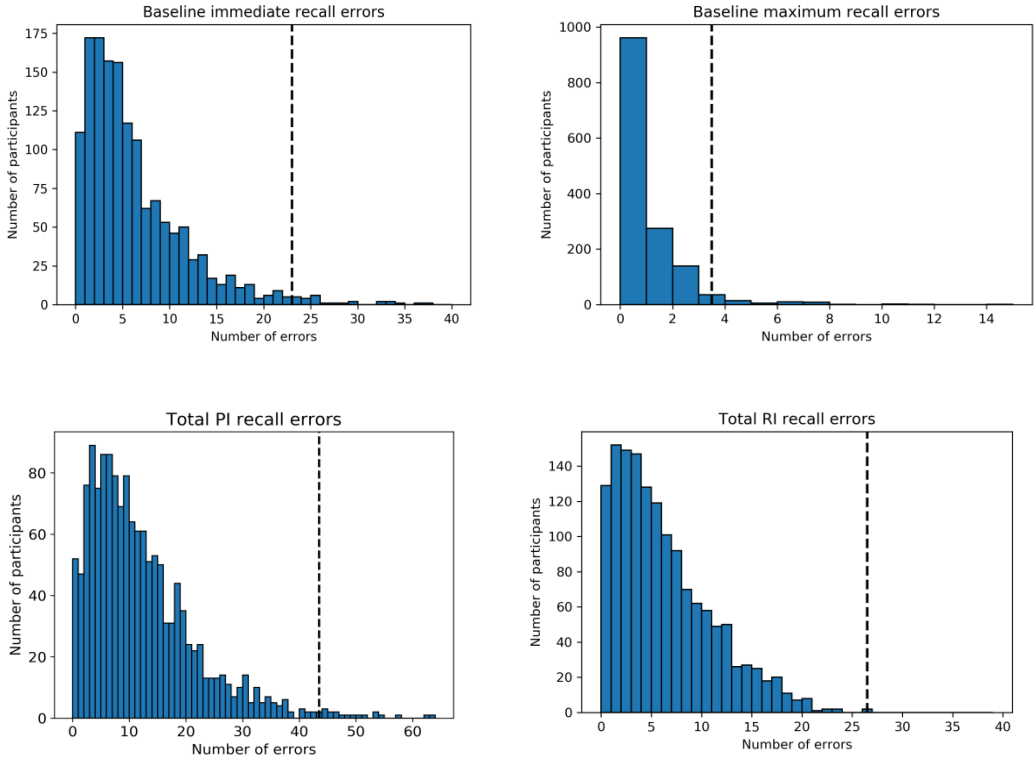


Figure 6.8. *Distributions of errors made by participants on the outcome variables.*

6.3 Statistical Analysis

To test the hypotheses central to this chapter, a series of Wilcoxon matched pairs signed rank tests and generalised linear mixed effects model analyses were

conducted. Specifically, the Wilcoxon matched pairs signed rank tests were used to assess whether the computationally generated confusable object pairs inflicted interference when used in Gallery Blast by testing for the differences in the number of recall errors made by the participants before and after the presentation of the confusable counterparts. In the PI condition, this was done by testing whether total PI recall errors made by the participants were significantly more than the baseline immediate recall errors. Since both these variables recorded immediate recall errors by the participants, their comparison would inform whether the computationally generated confusable object pairs inflicted PI. Similarly, in the RI condition, it was tested whether total RI recall errors were significantly more than the baseline maximum recall errors. Apart from calculating the p-values to check for significance, effect sizes (r) from these tests were also calculated. For reference, the interpretation values for r commonly found in published literature are: 0.10 - 0.3 (small effect), 0.30 - 0.5 (moderate effect) and >0.5 (large effect).

Similarly, to examine the relationships between the demographic characteristics and participant response on the task (in order to assess whether these align with existing literature), as well as to explore the specific effects of object pair similarity on interference, generalised linear mixed effects model analyses were conducted using the glmer function from the lme4 package in R (version 1.1.25, Bates et al., 2015). Linear mixed effect models yield parameter estimates of the predictor variables (e.g., demographic, and stimulus-level variables) that represent the size and direction of their relationships with the response variable. These models are desirable for use in Gallery Blast analysis for two key reasons. First, they control for the confounds when estimating the effects of variables of interest on the outcome measures. Second, by accounting for the random effects of participants and stimuli, they provide more robust parameter estimates of the predictor variables that can be extended to studies with different sets of participants and stimuli (Clark, 1973). To estimate the effects of the

variables of interest on PI, the PI intrusion error rates (see Table 6.2) were modelled using binomial regression on the proportion of intrusion errors made by the participants across trials in the PI condition. Binomial regression models estimate the odds of observing an event on a pre-determined number of trials, given a set of predictor variables. In the present analysis, this translates to estimating the odds of observing the number of intrusion errors on the number of recall trials attempted per stimulus per participant, given the set of demographic and stimulus-level predictor variables. Since the count of PI intrusion errors was bounded by a pre-determined number of trials, this model was the most appropriate in this setting. Similarly, the RI intrusion errors were modelled using logistic regression on participants' binary recall intrusion errors for a given stimulus in the RI condition ("no intrusion error" = 0, "intrusion error" = 1). Using these models, it was possible to estimate how the probability of making intrusion errors due to PI and RI were moderated by the demographic and stimulus-level variables.

To account for the random effects of participants as well as the stimuli used in Gallery Blast, "participant identifier" and "object name" were specified *a priori* as random intercepts in the linear mixed models. The variables in the fixed part of this analysis included stimulus-level (congruency and similarity ratings of object pairs), demographic-level (age, sex, education level, recorded family history of dementia and subjective cognitive function score) and theoretically determined interactions of age with baseline errors, family history of dementia, subjective cognitive function, and object pair similarity. Since AD is degenerative and its prevalence increases with advancing age, it is likely that factors such as family history of dementia and subjective cognitive function score might be more relevant in determining participant performance on Gallery Blast in the older cohort than in the younger cohort (see Khanahmadi et al., 2015 and Turner, 2006 for support). Similarly, age interaction with baseline performance would also highlight the differential effects of interference-free learning and recall on the interference-influenced performance on the task between older and

younger cohorts (Loewenstein et al., 2018). Finally, age interaction with object pair similarity would provide an insight into the change in the effect of object pair similarity on participant performance with age, thereby highlighting the potential use of this metric for sensitively recording cognitive differences between younger and older cohorts.

Before interpreting results from the respective models, necessary diagnostic tests were performed, and all the continuous predictor variables were standardized. Upon executing the models, the resulting coefficients (β) were used to calculate odds ratios (OR) as indices of the change in odds for a one-unit standard deviation (SD) change in the fixed effects of interest ($OR = e^{\beta}$). These have been used to report standardised effects of the predictor variables on the probability of committing intrusion errors. A positive β coefficient for a predictor variable indicates that the variable contributes to increased susceptibility to interference and vice versa. The significance level was set at p-values < 0.05.

6.4 Results

6.4.1 Computationally produced confusable pairs inflict interference effects

Significant differences were observed in the number of recall errors made by the participants before and after the presentation of the confusable counterparts in both PI and RI conditions. Specifically, in the PI condition, participants committed significantly more total PI recall errors (M = 11.65, SD = 9.49) than baseline immediate recall errors (M = 5.83, SD = 5.60) with a large effect size (Z_{1455} = 28.89; r=0.75; $p < 10^{-6}$). Similarly, in the RI condition, participants committed more total RI recall errors (M = 5.81, SD = 4.79) than baseline maximum recall errors (M = 0.68, SD = 1.36) with a large effect size as well (Z_{1455} = 32.93; r=0.86; $p < 10^{-6}$). Notably, RI effect size was larger than the PI effect size.

Table 6.3. Summary results from modelling PI and RI intrusion errors.

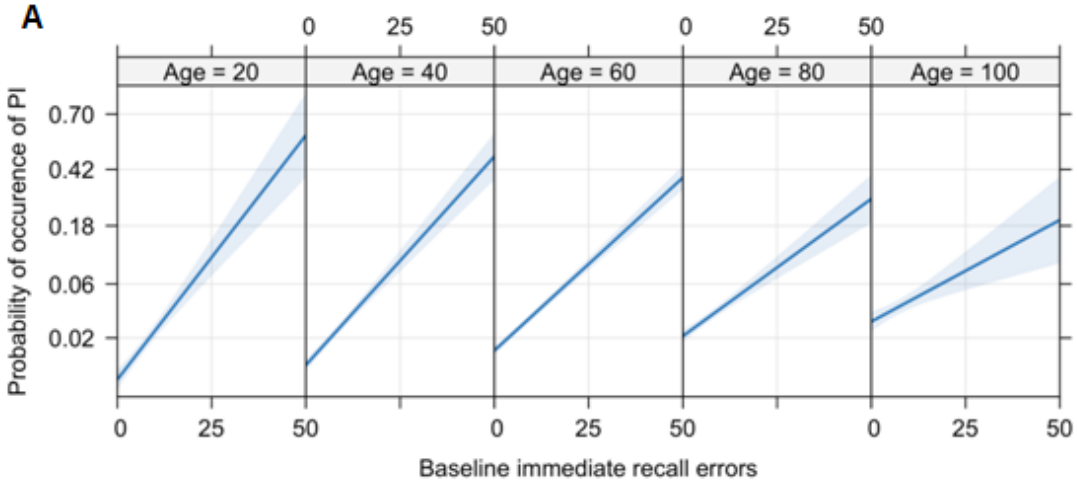
	PI effects					RI effects				
Variable	**Estimate**	**SE**	**z**	**CI (95%)**	***P***	**Estimate**	**SE**	**Z**	**CI (95%)**	***P***
(Intercept)	-3.668	0.081	-45.263	[-3.827, -3.510]	**< 0.001**	-1.749	0.164	-10.649	[-2.070, -1.427]	**< 0.001**
Object Pair Similarity	0.047	0.037	1.271	[-0.026, 0.121]	0.204	0.431	0.083	5.183	[0.268, 0.594]	**< 0.001**
Congruency [Same]	-0.245	0.061	-3.994	[-0.365, -0.125]	**< 0.001**	-0.452	0.135	-3.347	[-0.717, -0.187]	**0.001**
Congruency	0.072	0.055	1.309	[-0.036, 0.181]	0.19	-0.221	0.138	-1.605	[-0.492, 0.049]	0.109
Congruency [Perpendicular]	0.39	0.06	6.525	[0.273, 0.507]	**< 0.001**	0.205	0.134	1.53	[-0.058, 0.469]	0.126
Age	0.142	0.021	6.658	[0.100, 0.184]	**< 0.001**	0.094	0.033	2.832	[0.029, 0.159]	**0.005**
Baseline errors	0.42	0.016	25.963	[0.389, 0.452]	**< 0.001**	0.442	0.028	16.079	[0.389, 0.496]	**< 0.001**
Maximum education level	-0.029	0.013	-2.272	[-0.054, -0.004]	**0.023**	-0.129	0.02	-6.536	[-0.168, -0.090]	**< 0.001**
Biological Sex [MALE]	-0.044	0.043	-1.032	[-0.128, 0.040]	0.302	0.22	0.065	3.365	[0.092, 0.347]	**0.001**
Family history of Dementia	-0.071	0.034	-2.091	[-0.138, -0.004]	0.051	-0.063	0.053	-1.192	[-0.166, 0.041]	0.233
Subjective cognitive function	0.014	0.017	0.839	[-0.019, 0.047]	0.401	0.027	0.026	1.028	[-0.024, 0.077]	0.304
Age x Baseline errors	-0.05	0.016	-3.079	[-0.082, -0.018]	**0.002**	-0.096	0.028	-3.47	[-0.151, -0.042]	**0.001**
Age x Family history of Dementia	-0.063	0.036	-1.733	[-0.133, 0.008]	0.083	-0.023	0.057	-0.408	[-0.134, 0.088]	0.683
Age x Subjective cognitive function	0.008	0.018	0.439	[-0.028, 0.044]	0.661	-0.003	0.029	-0.12	[-0.060, 0.053]	0.905
Age x Object pair similarity	0.001	0.01	0.07	[-0.018, 0.019]	0.944	0.043	0.016	2.747	[0.012, 0.073]	**0.006**

6.4.2 Susceptibility to both PI and RI increases with age

Significant positive main effects of age were observed on susceptibility to both PI (β = 0.142, OR = 1.152, p < 0.001) and RI (β = 0.094, OR = 1.10, p = 0.005), indicating that susceptibility to both PI and RI worsened as a function of higher age. Specifically, a one standard deviation increase in age resulted in a 15.2% increase in the odds of occurrence of PI and a 10% increase in the odds of occurrence of RI. Notably, in both PI and RI conditions, the effect of an increase in age on task performance was found to be similar.

6.4.3 The effect of initial learning on errors due to both PI and RI decreases with age

Committing more errors respectively on interference-free baseline learning and recall due to global memory/attention deficit had significant positive main effects on susceptibility to both PI (β = 0.42, OR = 1.52, p < 0.001) and RI (β = 0.442, OR = 1.56, p < 0.001). However, a significant interaction with age, driven by a decrease in the effect of initial learning on errors due to both PI (β = -0.05, OR = 0.95, p = 0.002) and RI (β = -0.096, OR = 0.91, p = 0.001) with age was found. This is evident in Figure 6.6, where an increase in age prompts a reduction in the effect of baseline recall errors on both PI and RI.



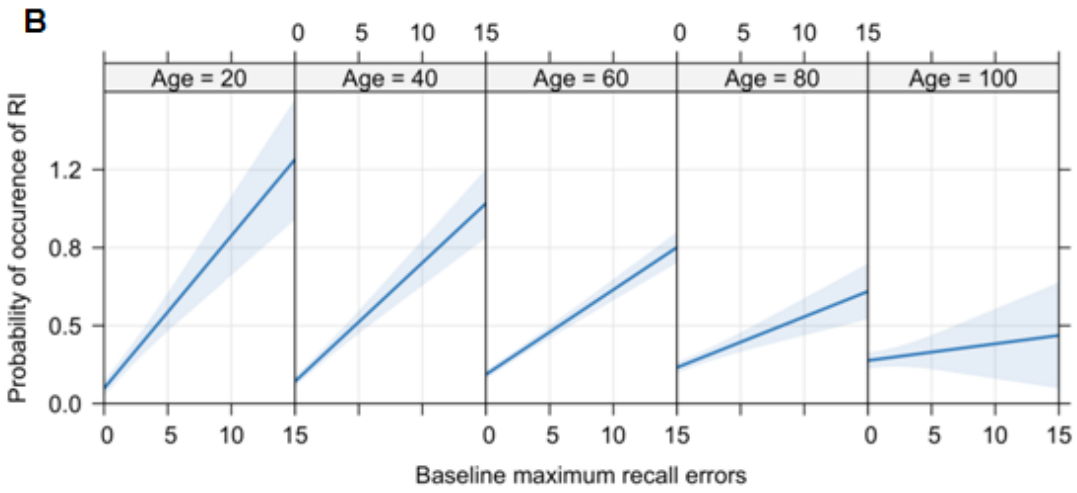


Figure 6.9. *Interaction effect of age and baseline errors on susceptibility to (A) PI, and (B) RI.*

6.4.4 Susceptibility to both PI and RI decreases with higher maximum education attained

Significant negative main effects of maximum education attained were observed on susceptibility to both PI (β = -0.029, OR = 0.97, p = 0.02) and RI (β = -0.129, OR = 0.879, p < 0.001), indicating that participants with higher maximum education were less susceptible to both PI and RI.

6.4.5 Susceptibility to RI but not PI increases with increase in object pair similarity

While a significant positive effect of object pair similarity was observed on susceptibility to RI (β = 0.431, OR = 1.54, p < 0.001), indicating that the participants found it harder to inhibit interference as the matched objects became more similar, the effect of object pair similarity was insignificant in the PI condition (p = 0.204), indicating that even with an increase in object pair similarity, there was no difference in susceptibility to PI among the participants (see Figure 6.7).

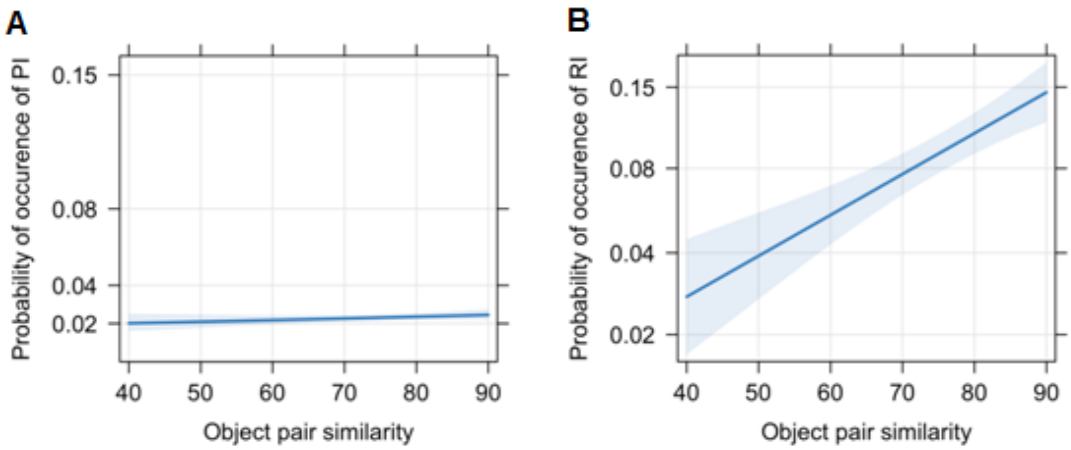


Figure 6.10. *Effect of object pair similarity on susceptibility to (A) PI, and (B) RI.*

6.4.6 Effect of object pair similarity increases with age in the RI but not the PI condition

A significant positive interaction between object pair similarity and age in the RI condition (β = 0.043, OR = 1.04, p = 0.006) indicated that with an increase in participant age, object pair similarity had a larger effect on the participant's ability to inhibit RI. A significant interaction effect, however, was not found between object pair similarity and age in the PI condition (p = 0.944).

To further explore the differential effects of object pair similarity on interference in different age groups, the participants were broadly divided into young and old age groups, where participants aged below 50 were classified as young, while participants aged above 65 were classified as old. These age ranges for the two groups were carefully selected after reviewing the norm in similar studies (see Khanahmadi et al., 2015 for the older and Loprinzi et al., 2018 for the young group). Following this, the mixed effects models were retrained on these sub-population data while controlling for all the variables controlled for in the main models. The differential effects of object pair similarity yielded by these models in the PI and RI conditions in the chosen age groups are shown in Figure 6.8. Specifically, the object pair similarity metric had a significant

and larger effect on the older age group's susceptibility to RI (β = 0.50, OR = 1.65, p < 0.001) compared to the younger group's susceptibility to RI (β = 0.36, OR = 1.43, p < 0.001). In the PI condition, while the magnitude of the effect of the metric followed the same trend as in the RI condition, i.e., higher for the older group, it was not significant in either age group (see Figure 6.8).

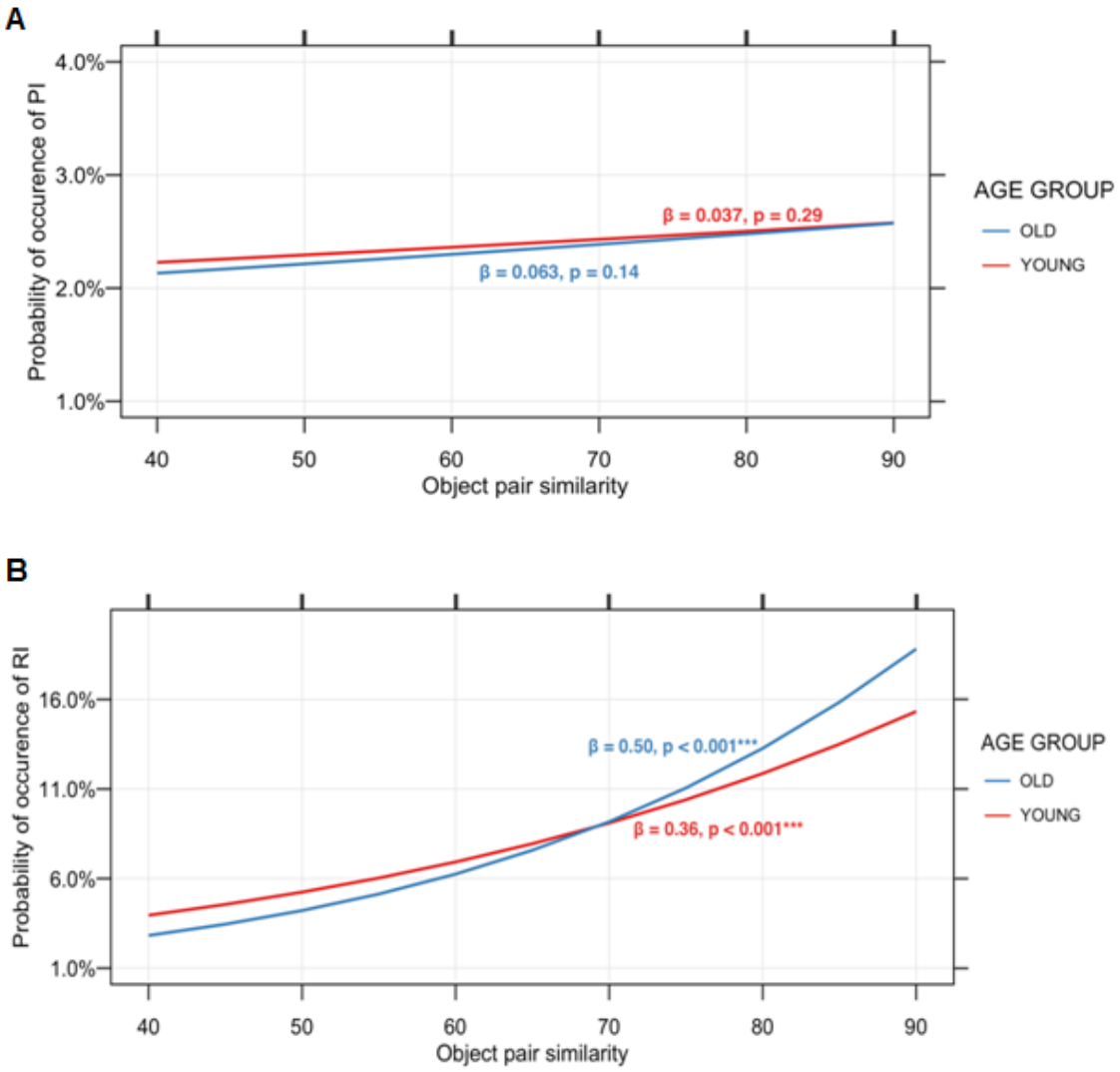


Figure 6.11. *Differential progression of the effects of object pair similarity from young to old in (A) PI and (B) RI.*

6.4.7 Stimulus-level covariate of no interest

The stimulus-level categorical covariate, congruency, was found to follow the expected pattern across levels in both the PI and RI conditions. Specifically, it was expected that when the direction associated with the object is perpendicular to the orientation of the

object, the participants get the least assistance to form strategies and are hence the most susceptible to committing errors. On the other hand, when the direction associated with the object is the same as the orientation of the object, the participants get the most assistance to form strategies and are hence least susceptible to make errors. As can be seen from the estimates of this variable in Table 6.3 and plotted in Figure 6.9, the probability of occurrence of both PI and RI follows the pattern: Perpendicular > Same with None and Opposite conditions following similar patterns. Despite this useful result, the utility of this variable is limited to controlling for the confounds of congruency in strategy formation in order to measure the true effects of other variables of interest.

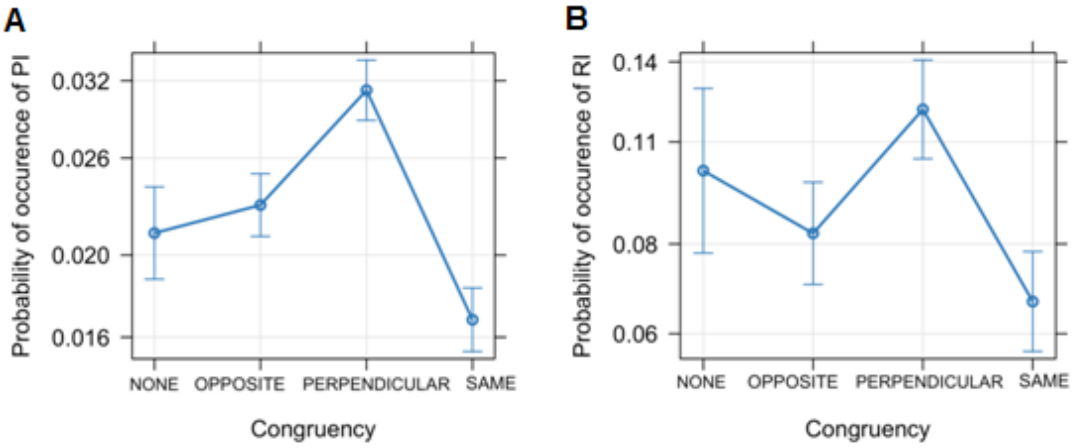


Figure 6.12. *Effects of different levels of congruency on susceptibility to (A) PI and (B) RI.*

6.4.8 Other demographic-level covariates

Significant effects on susceptibility to neither PI nor RI were found for biological sex, family history of dementia, subjective cognitive function, or their respective interactions with age, with the exception of male participants committing significantly more errors in the RI condition ($\beta = 0.22$, OR = 1.25, $p = 0.001$). The interpretation of these results is discussed in detail in the next section.

6.5 Discussion

In this chapter, I presented an investigation into the usability of the computationally generated confusable object pairs in an objective cognitive test, the Mezurio Gallery Blast. To conduct this investigation, four key aims were listed at the beginning of this chapter: 1) to examine the validity of Mezurio Gallery Blast as a reliable test of cognition; 2) to assess whether the computationally generated confusable object-image pairs can inflict interference in short-term memory when deployed in the task; 3) to assess the effects of object pair similarity on PI and RI; and 4) to assess the differential effects of the object pair similarity metric on healthy adults belonging to different age groups to examine the potential use of this metric for sensitively recording individual cognitive differences in the Gallery Blast task. Next, the findings from this investigation are discussed in detail.

6.5.1 Examination of the validity of the Gallery Blast task

To examine the validity of the Gallery Blast task as a test of cognition, performance across demographic variables was compared to findings within literature. The significant main effects of demographic-level variables in both PI and RI conditions in Gallery Blast showed that: 1) older age was associated with a higher number of intrusion errors, showing a reduced interference control ability in Mezurio Gallery Blast in older adults; 2) baseline performance in the no interference condition significantly determined the intrusion errors committed due to interference, indicating that reduced attention at the time of learning object-direction associations in Gallery Blast influenced susceptibility to interference on the task; and 3) adults with higher levels of education showed better interference control ability. As discussed in detail next, these results align with previous evidence of the relationships between the noted variables and interference in the working memory, thereby providing support to the functioning of the Gallery Blast task as well as the chosen outcome measures.

6.5.1.1 Relationship between interference and age

The main effects of age on susceptibility to both PI and RI reflected an increase in susceptibility to interference with an increase in age. These results extend a large body of previous evidence (Jonides et al., 2000; Loosli, Rahm, Unterrainer, Weiller, and Kaller, 2014; Samrani et al., 2017; Schmiedek et al., 2009; Stark et al., 2015; Vieweg, et al., 2015; Naveh-Benjamin and Mayr, 2018; Güsten et al., 2021) for an age-related reduction in the ability to control interference from competing stimuli in the working memory in healthy participants (see Samrani and Persson, 2021 for a review). As reviewed in Chapter 2, studies using other visual stimuli have shown the same pattern for healthy subjects where interference from similar stimuli impacted older participants more than younger participants (Schacter et al., 1992; Grady, 1995; Toner et al., 2009; Holden et al., 2012, 2013; Reagh et al., 2014; Sheppard et al., 2016).

Interestingly, a significant interaction between age and baseline recall performance on the task in both PI and RI conditions indicated that as age increases, the influence of baseline performance in explaining susceptibility to interference decreases. This result demonstrates that in older adults, the interference-free learning and recall performance is not as influential a determiner of their susceptibility to interference, as it is in younger participants. This means that while the susceptibility to interference experienced by the younger adults can be explained largely by poor attention or memory at the time of learning, susceptibility to interference in older adults needs an explanation beyond this general memory deficit (Crocco et al., 2014).

The inhibition-deficit theory (Hasher and Zacks, 1988) offers a potential explanation for this result. This theory suggests that while poorer quality of initial encoding and a "deficit" in the capacity to hold new memories significantly explains errors at recall, it is the ability to "inhibit" competing information at recall, a core function of the working memory, that declines significantly with age. Older adults are worse at inhibiting

competing memories than younger adults and it is this inability that primarily explains the errors made by older adults on cognitive tasks testing the ability to control interference. Correspondingly, inhibition is not severely impacted in younger adults and the errors they make on the interference tasks can be primarily explained by their baseline learning/recall performance. This theory further supports that interference from competing information, which is central to the design of the Gallery Blast task, and not mere baseline capacity, is necessary to distinguish between the cognitive differences in young and old healthy adults.

In fact, a non-significant correlation between age and baseline errors (Spearman's ρ = 0.05, p = 0.15) from participant performance on the Gallery Blast task was found that shows that the results from interference-free trials would not be sufficient to disambiguate between younger and older participants on the Gallery Blast task.

6.5.1.2 Relationship between interference and other demographic-level variables

Maximum education level was another demographic-level variable that significantly predicted the susceptibility to interference. A negative correlation with susceptibility to both PI and RI is consistent with a large body of literature that posits that adults with higher levels of education are buffered from susceptibility to interference (Hultsch et al., 1984; Zelinski and Gilewski, 1988; Harvey et al., 2017). Such studies have shown that high levels of education enable the participants to develop efficient strategies for encoding and processing of stimuli, thereby influencing their performance on such cognitive tasks (Harvey et al., 2017).

While the variables discussed thus far significantly influenced outcome measures in both PI and RI conditions, other variables that did not, also deserve some

commentary. These are biological sex, family history of dementia, subjective cognitive function, and their respective interactions with age.

To detail on the effect of biological sex on susceptibility to interference, a number of studies have historically reported no significant effect of biological sex on susceptibility to interference (MacLeod, 1991; Nigg et al., 2002; Rucklidge and Tannock, 2002; Mourik, et al., 2005; Dubuc et al., 2020), which aligns with the findings obtained for the non-significant effect of biological sex in the PI condition. However, a handful of studies have indeed shown gender-based differences with females outperforming males in neuropsychological tests (Naglieri and Rojahn, 2001; Mefoh, 2010; Spielberg et al., 2015). Such studies have posited that gender differences in brain structure could potentially explain the gender differences in cognitive control. For instance, Mefoh (2010) and Loprinzi and Frith (2018) attributed superior recall ability of women on interference tasks to the lateralisation effect (McGlone, 1978; Maitland et al., 2004; Krueger and Salthouse, 2010). It is important to note here that findings from the mixed effects model (Table 6.3) indicated that males performed significantly worse in the RI condition and therefore biological sex determined a participant's susceptibility to RI (on average) in the Gallery Blast task. However, to determine whether the lateralisation effect or other biological sex-related differences in cognitive control cited here are responsible for this finding is beyond the scope of this thesis but could be explored further in future studies.

Subjective cognitive decline (SCD) is another factor that has been widely explored as an indicator of incipient AD (Gauthier et al., 2006; Petersen et al., 1999). Studies have previously found SCD to be associated with an increased probability of amyloid positivity (Perrotin et al., 2012), white matter lesions (Minett et al., 2005), temporal lobe atrophy (von Gunten and Ron, 2004) and more general AD pathology (Barnes et al., 2006), thus supporting the validity of this marker in early detection of AD. In fact,

clinicians have historically primarily relied on subjective cognitive complaints to assess their patients for MCI or dementia. However, prominent studies and reviews (Zelinski et al., 2001; Purser et al., 2006; Mark and Sitskoorn, 2013) have argued that despite pathological support for SCD, it is not imperative that a relationship between SCD and objective cognitive tests can always be found. This could be due to a number of reasons. First, while cognitive decline in the healthy older population is modest at best and hence harder to measure subjectively, those with AD may estimate their cognitive functioning wrongly due to lack of self-awareness. Consequently, it is harder to form significant relationships with neuropsychological tests in both healthy and impaired older adults since these tests are more objective and possibly more sensitive to incipient AD. Second, patients' self-reports of cognition are usually impacted by their current state and could be heavily biased due to stressful life events in older life including but not limited to sickness, instability in life circumstances and widowhood, which might not directly reflect in their performance on objective tests (Wolf et al., 2005). In fact, Pavisic and colleagues (2021) have recently shown that the anxiety to develop dementia heavily influences self-reports of cognition in older adults. Finally, multiple studies suggest that SCD is linked more with future than current cognitive decline (see Reid and MacLullich, 2006 for a review) and may not always align with the current performance of participants on objective tests. It is proposed by these studies that the link between subjective cognition and performance on objective tasks is likely to be better established in longitudinal, rather than cross-sectional studies. As observed in the Gallery Blast analysis, a significant relationship was not observed between subjective complaint and performance on the task. It is perhaps possible, however, that longitudinal reports of cognitive decline over the duration of the GameChanger study may align with the corresponding longitudinal performance of the participants on Gallery Blast. Such a relationship would be interesting to look at as part of future work with Gallery Blast.

The final variables for which significant effects on participant performance in Gallery Blast were not found were family history of dementia and its interaction with age. While it is widely accepted that participants with a first-degree family history are at an increased risk of developing dementia (Turner, 2006), it remains unclear whether nondemented people with a positive family history of AD are more likely to experience cognitive deficits (Haussmann et al., 2018; Khanahmadi et al., 2015). It is not, therefore, always possible to obtain a significant relationship between family history and objective tests of cognition. Previously, Hayden and colleagues (2009), for instance, have shown that such a relationship is not well recorded even with the modified Mini-Mental state exam. Similarly, Morrow et al. (2009) have shown that family history of dementia alone does not significantly differ in scores on cognitive tests of memory, executive function, spatial ability, and attention. They did, although, find a significant interaction effect of family history with medical co-morbidities on subject performance on these cognitive tests and suggested that perhaps studies should collect information on medical co-morbidities among their participants. In another study, Edland et al. (1996) showed that maternal inheritance of AD is more likely than paternal, suggesting that perhaps questions on self-reports of family history of dementia should follow up with the question of lineage. Lastly, Donix et al. (2012) argue that perhaps a “yes” response to family history of dementia from the participants could be less specific for AD and could include genetic factors ranging from vascular and inflammatory processes affecting cognition or nongenetic risk factors, such as socioeconomic status or dietary habits affecting cognition. Their suggestion is to ask more for more details to confirm family history of AD status. Perhaps incorporating some of these suggestions in the future studies could show significant relationships between such factors and participant performance on objective cognitive tasks. Additionally, as part of future work, like with SCD, relationships between longitudinal participant performance on Gallery Blast and participant family history status could be evaluated.

To summarise this section, where significant, the results align with existing literature, thereby supporting the validity of Mezurio Gallery Blast as an accordant test of cognition. For the variables for which significant effects were not found, the findings were explained with literature support. Future work will include an evaluation of the potential relationships between longitudinal participant performance on Gallery Blast and their family history of dementia and SCD, or the interactions thereof.

6.5.2 Potential of the computationally generated confusable object pairs in inducing interference in Gallery Blast

To examine whether the computationally produced confusable object pairs interfered with participant memory, I tested whether the participants committed significantly more errors in retrieving the associated directions of the confusable matched counterparts of the previously learned objects. A higher number of total recall errors in both PI and RI conditions in comparison to their pre-interference equivalents showed that the participants did indeed experience interference from the computationally generated similar objects in Gallery Blast. These results align with the existing studies that have shown both PI and RI effects experienced by their participants (Loewenstein et al., 2004; Curiel et al., 2013; Crocco et al., 2014). Additionally, as discussed at the beginning of this chapter, there is a general agreement that RI effects are usually larger than PI effects. The findings in this chapter align with this agreement since in the Gallery Blast analysis too, RI effects on the working memory were higher than PI effects for the participants. Even more importantly, that the interference effects were found to exist across the stimulus set presented to the participants shows that not just a small sample, but the entire set of stimulus pairs were confusable enough to generate interference effects in participant memory.

6.5.3 Differential effects of the object pair similarity metric on PI and RI

A central aim of this thesis was to examine the potential role of object pair similarity in sensitively recording individual cognitive differences among the participants. For this, first the differential relationship between the object pair similarity metric and intrusion errors due to PI and RI was measured while controlling for the available demographic-level and stimulus-level variables. The results from the Gallery Blast analysis showed that object pair similarity significantly influenced susceptibility to RI (OR = 1.54; $p < 0.001$) but not to PI ($p > 0.2$). In other words, while the relationship between object pair similarity and susceptibility to RI aligned with the literature presented in Chapter 2, a lack of significance of the effects of object pair similarity on susceptibility to PI shows that the participants experienced similar interference effects irrespective of a change in object pair similarity.

While much of earlier research abandoned this problem of differential relationship between stimulus pair similarity and the two interference conditions by assuming the relationship between RI and stimulus pair similarity as “straightforward” and that between PI and stimulus pair similarity as “complicated” (Keppel and Underwood, 1962), recent research (Jonides et al., 2008; Sosic-Vasic et al., 2018; Chanales et al., 2019; Antony and Bennion, 2020) suggests that the differential effects of object pair similarity on PI and RI can be explained by how target memories differentially experience interference in these two conditions. Specifically, while in the RI condition, interference from a similar stimulus occurs only after the encoding process for the target stimulus is complete (where the encoding process is entirely interference-free), in the PI condition, interference occurs during the encoding of the target stimulus, since the interfering stimulus has already been presented before the imminent encoding of the target stimulus. Consequently, in the RI condition, the interfering stimulus competes for retrieval with a highly consolidated target stimulus in short-term memory – a competition that has been consistently shown in previous literature to

increase with an increase in content similarity (see studies cited in section 6.1). On the other hand, in the PI condition, the reactivation of the old interfering memory during the encoding of the target stimulus has the opposite effect, i.e., this reactivation of old similar memory has been shown to form strong links between the old and target memories (van Kesteren et al., 2018), thereby sometimes resulting in a relationship that counteracts competition from similar memories with cooperation among these memories (Kuhl et al., 2010; Koen and Rugg, 2016). In other words, in the PI condition, at certain values of similarity between the similar memories, the matched counterpart helps consolidate the memory of the target stimulus instead of weakening it as in the RI condition. This cooperation effect has been shown to be dominant in high similarity conditions (Chanales et al., 2019; Antony and Bennion, 2020).

The differential mixing of the cooperative and competing effects at different levels of similarity means that interference is substantially reduced (although not necessarily eliminated) in the high similarity condition and could be the reason that the relationship between stimulus pair similarity and PI effects is termed “complicated” in previous literature. While answering questions on the underlying working of short-term memory is beyond the scope of this thesis, this theory could potentially explain why a non-significant effect of object pair similarity was witnessed in the PI condition in Gallery Blast, viz. the interference effects could have been suppressed in the high similarity condition due to the cooperation effect, while remaining high in the low similarity condition, thereby reducing the slope of the effect of object pair similarity on susceptibility to PI compared to RI.

6.5.4 Differential effects of the object pair similarity metric on different age groups

Finally, to measure the differential effects of object pair similarity on interference in different demographic groups, the interaction effects between age and object pair similarity were obtained. As discussed earlier, there is ample evidence from the main analysis to suggest that susceptibility to both PI and RI in Gallery Blast increased with an increase in age. It was, therefore, justified to assess the interaction between object pair similarity and age for this purpose. The interaction effects from the main model revealed that while the effect of object pair similarity increases with increasing age in both the PI and RI conditions, such an increase was only significant in the RI condition.

For further exploration and with the intention of assessing the differential effects of object pair similarity on interference in different age groups, this effect was examined separately in sub-groups of young and old adults in both conditions. While interpretational limitations are inherent to analysing trends over categorical variables, retraining the mixed effects models on the two age-groups in the PI and RI conditions revealed that: 1) the effects of object pair similarity continued to be non-significant for both the age groups in the PI condition, 2) the effects of object pair similarity were significant for both the age groups in the RI condition and 3) increasing object pair similarity by one standard deviation has approximately a 50% higher average effect on older participants (OR = 1.65) than on younger participants (OR = 1.43) in the RI condition. Of note, the susceptibility to RI increases non-linearly with object pair similarity such that the increase in the effect of the metric is steeper at higher values of object pair similarity. In sum, assessing the differential impact of object pair similarity on the younger and older sub-cohorts has demonstrated the ability of this metric to sensitively record group differences in the RI condition in Gallery Blast.

6.5.5 Potential utility of the object pair similarity metric in detection of impairment

Similar to the work presented in this chapter, a large body of previous literature (see review in Chapter 2) has used visual stimuli to measure the differential effects of stimulus pair similarity on interference in young and aged adults. These studies had the availability of classification labels of the older participants as belonging to healthy or impaired sub-groups based on participant performance on standard neuropsychological tests or biomarker evidence, and generally found that the effects of stimulus pair similarity on task performance follow a progressive trajectory from young healthy adults to older healthy adults to older impaired adults as shown in one such example from Reagh et al. (2014) in Figure 6.10. This theory of progressive effects is commonly accepted by a large number of authors, many of whom have questioned the qualitative difference in performance between healthy and pathological ageing adults (Fotuhi, et al., 2009), observing that albeit to a lesser extent, the deficits observed in aMCI are also present in healthy aging (Buckner, 2004). Following this theory, it is likely that in future studies with high-risk cohort, performance on Gallery Blast follows a similar trajectory, in which case, it can be proposed that: 1) the performance of impaired old adults on Gallery Blast could be more sensitive to change with stimulus pair similarity in the RI condition as compared to the PI condition and 2) impaired old adults could face stronger effects of object pair similarity than both young and unimpaired old adults in the RI condition in Gallery Blast. Thus, increasing the similarity in the similarity range of stimulus pairs tested in this chapter may enhance the detection of cognitive differences associated with very early AD in Gallery Blast.

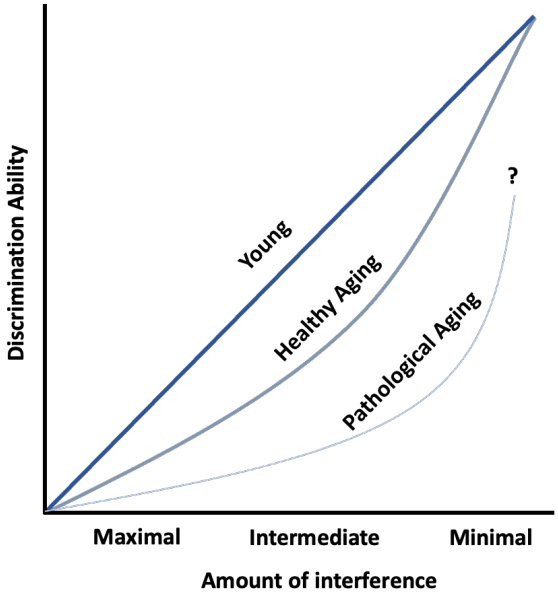


Figure 6.13. *Progressive effect of interference postulated from young and healthy to pathological aging (adapted from Reagh et al., 2014). Please note, where authors use "amount of interference" present thesis uses degree of confusability.*

Although such conclusions are speculative given the available data, this finding is promising for future use of Gallery Blast in the detection of preclinical AD. Future work, however, should seek to validate this prediction based on Gallery Blast outcomes against cerebrospinal fluid, positron emission tomography and magnetic resonance biomarkers indicative of AD pathology and subsequent cognitive impairment. Alternately, correlating Gallery Blast outcomes with participant performance on standard in-clinic tests of working memory might also provide the first indication of whether the proposed theory of higher sensitivity to an increase in object pair similarity in the RI condition is true.

Before concluding this chapter, it is imperative to address the expected effect of object pair similarity in high-risk cohorts in the PI condition. Two possibilities could emerge here. First, it could be possible that like the older group, the impaired group experiences similar interference across the similarity range (albeit making more errors than the older group), in which case, participant performance in the PI condition will not be sensitive to change with changing object pair similarity. However, in another possibility, high-risk cohort performance could vary across the object pair similarity

range as explained next. As seen in Figure 6.8A, the p-value for the effect of object pair similarity on the older cohort ($p = 0.14$) in the PI condition is lower than the p-value for the effect in younger cohort ($p = 0.29$). If the significance of the effect of object pair similarity continues to increase (i.e., the p-value continues to decrease) along with the effect size of the object pair similarity metric from the younger to the older to the impaired cohorts, the effect of the similarity metric in the PI condition, too, may become significant and higher than that for the older cohort. This could allow for differentiating between healthy and impaired older adults in the PI condition, viz. the performance of impaired old adults on Gallery Blast could be sensitive to change with object pair similarity in the PI condition too. However, as discussed above, future work will be needed to validate this prediction by deploying Gallery Blast in studies with high-risk cohorts.

6.6 Limitations

Although this research is a critical first step in establishing the scientific value of Mezurio Gallery Blast and the computationally matched confusable object pairs ahead of more costly clinical validation work, there are limitations worthy of note. These are as follows:

1. A major goal of this chapter was to evaluate interference effects from similar objects with overlapping features in accordance with the AB/A'C paradigm. However, many studies utilising this paradigm also use control pairs DE instead of AA' (thus making the paradigm DB/EC), where DE are completely unrelated pairs. This is done in order to record the effects of interference in the absence of similar memories. A few such trials in Gallery Blast would provide even better estimates of the effects of similar stimuli on interference in short-term memory.

2. Another shortcoming of the analysis in this chapter is that brain data were not available for evaluating the outcomes from Gallery Blast. Ongoing work seeks to validate Gallery Blast outcomes against cerebrospinal fluid, positron emission tomography and magnetic resonance biomarkers indicative of AD pathology and subsequent cognitive impairment.
3. The individuals included in this research have obtained high levels of education which may influence the generalizability of current findings. While adjustment for the education in the analyses presented may have made the findings more generalisable, a future study with participants without high levels of education may help in validating the results from the current analyses.
4. While the study used object pairs that were conceptually and visually dissimilar (e.g., <*lion, tiger*>), a comparison with similar studies that have used conceptually the same but visually dissimilar object pairs (e.g., <*cat1, cat2>)* was not done, which leaves a gap in research about which type of object pairs have more potential to differentially affect healthy and vulnerable populations.
5. While the potential utility of the object pair similarity metric in detection of impairment has been discussed in section 6.5.5, it needs more support to show that this metric can indeed differentially affect the healthy and the impaired older individuals as it does the younger and the older participants.
6. Finally, while this analysis presents a thorough investigation into the interference in short-term memory inspired blocks in Mezurio Gallery Blast, the randomized recall block at the end of the task (see figure 6.3) was left unexplored due to a lack of existing literature and hence a source to draw hypotheses on the topic. Upon the availability of biomarker evidence or the results from established neurocognitive assessments, results from these blocks should, however, be explored.

6.7 Conclusion

The present study demonstrates the usability of computationally matched similar object pairs in an objective test of cognition targeting short-term memory. The study also provides preliminary validation of Gallery Blast, a newly developed smartphone task, as a valid test of cognition in a remotely conducted large-scale study with participants from varied demographics. The data from the study support the value of incorporating confusable object pairs to capture interference effects in short-term memory. An age-related decline in inhibiting interference effects (both proactive and retroactive) from similar memories aligns with similar results found in previous literature and supports the future validation of interference-related outcomes from Gallery Blast against biomarkers of degeneration. The differential effects of object pair similarity in younger and older cohorts support increasing the similarity between old and interfering object stimuli in the RI condition to increase measurement sensitivity in detecting performance differences in healthy adults. These results are promising in establishing the use of Gallery Blast with the incorporated stimuli in sensitively recording group differences and may also facilitate the detection of AD at an earlier stage of progression.

Chapter 7

Conclusion

The overall aim of this thesis was **to explore the potential of computational models of vision and semantics, as well as their combinations, in generating datasets of parametrically varying confusable object image pairs usable in memory assessment tasks and to examine the utility of these computationally matched object pairs in an objective test of short-term memory**. To conduct this investigation systematically, the following steps were followed. First, the computational models of vision and semantics chosen for investigation in this thesis were introduced. Next, databases of unique object pairs parametrically varying in similarity were generated using similarities between objects captured by the computational models and their weighted combinations. The models and the generated datasets were then assessed on standard evaluation techniques. Following this assessment, human-rated similarities for the object pairs matched by the top-performing models were obtained. These similarity ratings were used to further validate the models and to investigate the usability of the computationally matched object pairs as well as their computationally generated similarity ratings in downstream applications such as memory assessment. After thorough investigation, these object pairs were deployed in a memory assessment task called Mezurio Gallery Blast, which was remotely attempted by a large sample of UK-based participants with varying demographics under the GameChanger study. Finally, following this study, the analysis of participant performance on the task was conducted to assess the utility of the computationally matched object pairs at stimulating confusability in short-term memory, as well as in sensitively recording individual cognitive differences, when deployed in the Gallery Blast task. In this chapter, I will revisit the key findings from this investigation, highlight the contribution of this thesis to the existing scientific knowledge and discuss future directions.

Computational models produced object pairs with varied degrees of similarity, but these object pairs needed validation from humans before use in a memory assessment task

The computational models used in this thesis produced object pairs that varied in perceptual similarity. A histogram of the number of human-rated object pairs distributed by mean similarity ratings from the validation study (Figure 5.6) showed that the rated object pairs were roughly uniformly distributed across similarity bins plotted with similarity indices ranging from 0 to 100. 880 of the 1,498 total image pairs generated by the top-3 best performing models were rated perceptually more similar than dissimilar (i.e., rated >=50) by the human participants. The availability of such object pairs varying in perceptual similarity has helped in establishing stimulus-level effects of object pair similarity on participant performance in Mezurio Gallery Blast. Furthermore, the computational models were able to create such datasets of confusable object pairs only by taking single object images and their names as inputs, without the need to retrain from scratch on large sets of images or text, thereby demonstrating the potential of "one-shot learning" at producing such datasets in a quick and computationally efficient manner.

However, further evaluation of the usability of the computationally generated object pair similarity ratings revealed that even the ratings from the best performing computational model were much less correlated than the average annotator's ratings to ImSim-1498. From this analysis, I concluded that even the best performing model employed in this thesis is some way from automatically producing object pair similarity ratings that can be trusted with the same confidence as their manually rated equivalents. As such, while these models give us the independence to automatically and quickly produce confusable object pairs varying in similarity merely from a dataset of single and isolated object images commonly used in psychiatric experiments, such object pairs must be validated for similarity using systematically designed human

rating experiments before using them for measuring stimulus-level effects in tests of memory and cognition. It is also worth mentioning here that since a large dataset of objects (consisting of 1,402 objects) was used to measure the performance of each computational model against the systematically acquired ImSim-1498 dataset of human-rated perceived similarity between objects, the performance rankings of these models at capturing the human knowledge of perceived object similarity obtained in this thesis (Chapter 5) can be trusted with high confidence. As such, I found that the visuo-taxonomic model integrating the similarity relationships captured by the visual and the taxonomic models outperformed all other models at estimating the human knowledge of perceived object similarity, highlighting that taxonomic input is a more reliable source of semantic information than linguistic input for the given application. As noted in Chapter 5, the ImSim-1498 dataset was useful in measuring the performance of the models at capturing the human knowledge of perceived object similarity, which is central to the application of these models in this thesis, since it is highly likely that this dataset captured truer estimates of overall "perceived" similarity between objects in comparison to the earlier word-based human-rated datasets, which might be more useful at measuring the performance of computational models in natural language processing-based applications such as document similarity as identified by Hill et al. (2015).

Computationally matched object pairs inflicted confusability in an objective test of cognition targeting short-term memory

When deployed in Mezurio Gallery Blast - an objective cognitive test targeting short-term memory, the computationally produced object pairs were found to generate significant interference effects in short-term memory across participants. A Wilcoxon matched pairs signed rank test conducted to assess whether the participants were able to inhibit competition at recall in short term memory from the newly presented computationally matched counterparts of the previously learned stimuli revealed that

the effect size of interference inflicted by such confusable counterparts was large in both the PI and the RI conditions. Furthermore, aligning with a large body of literature (see Schmeidler, 1939; McGeoch and Underwood, 1943; Melton and Von Lacrum, 1941; Underwood, 1945; Curiel et al., 2013; Crocco et al., 2014), the net interference effects were found to be higher in the RI condition than in the PI condition in Gallery Blast.

It is worth mentioning here that before finding the interference effects inflicted by the computationally matched pairs on Gallery Blast, it was concluded that the task demonstrated the properties of a valid test of cognition, which was hypothesised to be true if the relationships of the demographic-level variables with the outcome measures from the task would align with those widely accepted in literature. While this was an ambitious hypothesis, given a sizeable number of demographic-level variables recorded in the GameChanger study as well as the lack of agreement on the effects of many variables on the outcome measures in objective tests of cognition, it was found that where significant, the results aligned with existing literature. Specifically, the relationships demonstrated by demographic-level variables age, baseline performance and maximum education level with the outcome measure from Gallery Blast were significant and aligned with the those found in literature. Additionally, a significant interaction between age and performance on the baseline task demonstrated that merely assessing participant performance on baseline recall of the object-direction pairs would not have been sufficient in detecting group-level cognitive differences in short term memory and that the interference from competing stimuli was required to sensitively record such effects. It is also important to note here that significant effects of the demographic-level variables such as biological sex, family history of dementia and subjective cognitive decline and their interactions with age were not always found, but it was proposed that future work assessing the relationships between the year-on-year change in participant performance on the task and such variables could

potentially yield significant estimates of such variables. In sum, the findings summarised here supported the validity of Mezurio Gallery Blast as an accordant test of cognition and the significant differences in participant performance before and after the presentation of the confusable counterparts of previously learned stimuli demonstrated the utility of the computationally matched object pairs in inflicting confusability in an objective test of cognition targeting short-term memory.

Computationally produced object pairs parametrically varying in similarity allow for sensitive recording of susceptibility to interference in short-term memory

In the earliest chapters of this thesis, I illustrated the need for the production of datasets of confusable object pairs where the constituent objects were conceptually different. Specifically, I identified that either the confusable object pairs in the existing datasets consisted of visually distinct but conceptually the same objects (such as two visually distinct *cups*) or there was no record of similarities between the objects in pairs where both the constituent objects were visually and conceptually distinct. While the former did not allow for the testing of the effects of the object pairs used in this thesis for memory assessment, the latter did not allow for the fine-grained analysis of the effects of object pair similarity in memory assessment. However, as pointed out in earlier chapters of this thesis, such an analysis using stimulus pairs constituting conceptually distinct but similar objects was needed since 1) it is widely agreed that increased interference from such confusable pairs of stimuli can provide the earliest marker of preclinical AD and 2) parametrically varying stimulus pair similarity metric could provide a sensitive marker for recording individual cognitive differences, thereby assisting in early detection of neurodegeneration in adults affected significantly more than the baseline population on this metric. In the analysis conducted across the performance of participants in Gallery Blast, I found support for these hypotheses. Specifically, that the mean effect of object pair similarity in the RI condition significantly differed between the younger and older age groups who attempted the Gallery Blast

task (older age group affected more than young age group with increasing object pair similarity) clearly demonstrated that the object pair similarity metric, as employed in this thesis, sensitively records between-group cognitive differences. Moreover, as noted in Chapter 6, since a large body of literature such as that reviewed in Chapter 2 has found progressively worsening effects of stimulus pair similarity on task performance from young healthy adults to older healthy adults to older impaired adults, I hypothesise that future validation of Gallery Blast against biomarkers or established neuropsychological tests could show impaired older adults as the group experiencing the largest effects of increase in object pair similarity on their performance in Gallery Blast, especially in the RI condition.

Even apart from the support for the utility of the object pair similarity metric in early detection of neurodegeneration, the findings from this thesis are useful to the wider field of psychiatry as explained next. Recent studies such as those by Chanales et al. (2019) have shown that at the highest value of similarity between an old and a new memory, reactivation of the older memory during encoding of the new memory (PI condition) strengthens the two memories due to a cooperation effect between them (see section 6.5.3 for details), thereby reducing the errors on recall of such memories. To mark the similarity between the old and the new memory, the authors in the said study relied on the same/different paradigm, i.e., same memories were 100% similar, while different memories were 0% similar per their definition. Using this categorical value of similarity, they were able to demonstrate that when the similarity was 100%, the recall error was considerably lower in the PI condition compared to the RI condition in their experiment, while when the overlap was 0%, the recall error remained similar in the two conditions. In sum, they showed that at the highest value of similarity, fewer interference errors were made by the participants in the PI condition. However, this analysis did not demonstrate the effect of granular values of object pair similarity on recall errors. By carefully controlling the preciseness of object pair similarity at a

granular level and thereby obtaining the effects of this metric on interference in short-term memory on a continuous scale, I was able to extend the same/different paradigm of object pair similarity and contribute to the wider field by demonstrating the differential effects of parametrically varying object pair similarity on both PI and RI. Specifically, my findings suggest that the recall errors due to PI made by the participants were similar across the spectrum of object pair similarity used in this thesis, whereas the recall errors made due to RI increased with an increase in object pair similarity. Extending the findings obtained by Chanales et al. (2019), I have reasoned that while at lower values of object pair similarity, the interference in short-term memory remains comparable between PI and RI due to sufficient competition, with increasing values of similarity, the memories face competition as well as cooperation from their matched counterparts in the PI condition, thereby reducing the recall error rate at higher values of similarity in comparison to the RI condition. Given my findings, I hypothesise that with an increasing value of similarity in the PI condition, the cooperation effect between memories increases while the competition effect decreases. Since my findings are in a range of similarity (42-89 on a scale of 100) between those used by Chanales et al. (2019), it is possible that extending the similarity range from 0 to 100 may yield converging results with them, i.e., lower recall errors at 100% similarity than at lower values of similarity. In sum, this finding has the potential to extend our knowledge of the granular effect of object pair similarity in the PI condition. Next, I summarize the key contributions of this thesis.

7.1 Summary of contributions

1. Creation of a dataset for the Mezurio Gallery Blast task: A major contribution of this thesis is the creation of a dataset consisting of matched object pairs parametrically varying in similarity. As discussed in this thesis, this dataset has assisted in sensitively recording the behavioural differences in cognitively normal

adults belonging to different age groups, thereby providing preliminary evidence of the ability of the task to provide a signal in the very earliest stages of AD. Consequently, this dataset is now actively deployed as part of Mezurio Gallery Blast in the following studies to establish the utility of smartphone based digital tools in early detection of AD: 1) RADAR-AD (Owens et al., 2020) – currently active in over 10 European countries on participants with and without impairment; 2) the EDoN initiative (Frey et al., 2019), a global multi-cohort early detection project active and collecting participant performance data on smartphone-based digital biomarkers in the US, with plans to extend the study to the UK, Australia and many other countries; 3) the GameChanger study, as highlighted in this thesis and finally 4) in some other clinical studies such as the "Downregulating Functional Hyperactivity in APOE e4 carriers" study hosted at the University of Sussex.

2. Establishing the construct validity of the Mezurio Gallery Blast task: Another major contribution of the work undertaken in this thesis was establishing the construct validity of the Gallery Blast task, which is integral to the overall success of the Mezurio app platform. By demonstrating that the relationships between participant demographics and the identified outcome measures from the task align with that found in existing literature, the work in this thesis demonstrated that the task works as intended and sensitively captures individual differences in short-term memory function. This contribution establishes confidence in the task for the stakeholders involved in the studies discussed above where the task is currently or in the process of being deployed.
3. Demonstrating the value of datasets of parametrically varying confusable object pairs: The constituent objects in the confusable object image pairs used in this thesis are conceptually distinct, which means that they are not merely two different looking exemplars of the same concept (such as *cup-cup*) such as those used in earlier research reviewed in Chapter 2. Two novel contributions of this thesis,

therefore, are 1) demonstrating that such confusable object pairs too, inflict confusability in short-term memory across demographics, and 2) extending our knowledge of the relationship between interference in short-term memory and object pair similarity on the stimulus pair type previously untested in short-term memory assessment.

4. Adding to the current knowledge about the influence of parametrically increasing object pair similarity on PI vs. RI: A debate on the differential effects of increasing stimulus pair similarity on PI and RI can be found in literature from the early 20th century till today. The findings from this thesis add new knowledge to this debate by presenting the granular relationships between object image pair similarity and PI and RI. Specifically, the findings from this thesis show that while RI increases (non-linearly) with an increase in object image pair similarity, the magnitude of PI remains similar through the spectrum of similarity, possibly due to the opposing effects of competition and cooperation between memories at different values of similarity in this condition.
5. Establishing the value and limitations on the use of computational methods for creating datasets central to this thesis: Finally, this thesis also tested the capability of machine learning based computational methods at producing the datasets of parametrically varying confusable object pairs usable in memory assessment tasks. The present work showed that while the computational models used in this thesis are capable of producing object pairs with overlapping visual and semantic features, they are far from producing similarity ratings as reliable as those produced by an ensemble of humans. This thesis concludes that while such methods are useful for quickly producing datasets of confusable object pairs parametrically varying in similarity, the similarity ratings of the matched object pairs should be rated by humans using a carefully designed rating experiment before using such pairs in downstream applications such as the memory assessment task central to this thesis. Future studies could use more advanced

computational models of vision and semantics to produce confusable object image pairs using the approach suggested in this thesis.

7.2 Future directions

While the findings from this study align with and contribute to a large body of existing literature, I did not find significant effects of family history of dementia and subjective cognitive function on the outcomes from the Mezurio Gallery Blast. I propose that future work should assess the effects of such variables on a year-on-year change in participant task performance, which could yield a significant effect of such variables on task performance, especially in older adults.

Additionally, while the GameChanger study provided access to a large cohort of participants with varied demographics, which was essential for the analysis undertaken in this thesis, a lab-based study with select cohort is needed to assess whether the findings obtained in this thesis can be replicated in focussed cohorts. It would be even more beneficial if such a study could include participant performances on other established neuropsychological tests.

Finally, a critical task following this thesis is the validation of the outcomes from Mezurio Gallery Blast against established cognitive tests and cerebrospinal fluid, positron emission tomography, and magnetic resonance based biomarkers indicative of AD pathology and subsequent cognitive impairment.

7.3 Concluding remarks

This thesis introduced a novel method for obtaining confusable object pairs parametrically varying in similarity, where the objects in the pairs were both visually and conceptually different but were similar to some degree. Carefully designed cognitive tasks such as the Mezurio Gallery Blast have the potential to utilise such

pairs for the assessment of the health of the cortical regions in the brain that are first vulnerable to neurodegeneration. This thesis showed that such computationally matched object pairs are not only useful in generating interference effects when employed in an objective cognitive task but can also capture group-level differences in cognition, thereby providing preliminary support for the use of these pairs for measuring individual cognitive differences using the Gallery Blast task.

Appendix A: WordNet Synsets and definitions of semantic categories

	Concept	Synset	Wordnet definition
1	ACCESSORY	accessory.n.01	clothing that is worn or carried, but not part of your main clothing
2	APPLIANCE	appliance.n.01	a device or control that is very useful for a particular job
3	BIRD	bird.n.01	warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings
4	CLOTHING	clothing.n.01	a covering designed to be worn on a person's body
5	CRUSTACEAN	crustacean.n.01	any mainly aquatic arthropod usually having a segmented body and chitinous exoskeleton
6	ELECTRONICS	electrical_device.n.01	a device that produces or is powered by electricity
7	FISH	fish.n.01	any of various mostly cold-blooded aquatic vertebrates usually having scales and breathing through gills
8	FLOWER	flower.n.01	a plant cultivated for its blooms or blossoms
9	FOOD	food.n.02	the flesh of animals (including fishes and birds and snails) used as food
10	FRUIT	edible_fruit.n.01	edible reproductive body of a seed plant especially one having sweet flesh
11	FURNITURE	furniture.n.01	furnishings that make a room or other area ready for occupancy
12	HOUSEHOLD	instrumentality.n.03	an artifact (or system of artifacts) that is instrumental in accomplishing some end
13	INSECT	insect.n.01	small air-breathing arthropod
14	INSTRUMENT	instrument.n.01	a device that requires skill for proper use
15	MAMMAL	mammal.n.01	any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk
17	MOLLUSC	mollusc.n.01	invertebrate having a soft unsegmented body usually enclosed in a shell
18	MUSIC	instrument.n.06	any of various devices or contrivances that can be used to produce musical tones or sounds

19	PLANT	plant.n.02	(botany) a living organism lacking the power of locomotion
20	REPTILE	reptile.n.01	any cold-blooded vertebrate of the class Reptilia including tortoises, turtles, snakes, lizards, alligators, crocodiles, and extinct forms
21	RODENT	rodent.n.01	relatively small placental mammals having a single pair of constantly growing incisor teeth specialized for gnawing
22	SPORT	game_equipment.n.01	equipment or apparatus used in playing a game
23	STATIONERY	implement.n.01	instrumentation (a piece of equipment or tool) used to effect an end
24	STRUCTURE	structure.n.01	a thing constructed; a complex entity constructed of many parts
25	TOOL	tool.n.01	an implement used in the practice of a vocation
26	TRANSPORT	transport.n.01	something that serves as a means of transportation
27	UTENSIL	utensil.n.01	an implement for practical use (especially in a household)
28	VEGETABLE	vegetable.n.01	edible seeds or roots or stems or leaves or bulbs or tubers or nonsweet fruits of any of numerous herbaceous plant
29	WEAPON	weapon.n.01	any instrument or instrumentality used in fighting or hunting

Appendix B: List of all concepts and their WordNet Synsets

	Concept	Synset
1	aardvark	aardvark.n.01
2	abacus	abacus.n.02
3	accelerator	accelerator.n.01
4	accordion	accordion.n.01
5	acorn	acorn.n.01
6	agouti	agouti.n.01
7	air conditioner	air_conditioner.n.01
8	airplane	airplane.n.01
9	airship	airship.n.01
10	alarm	alarm.n.04
11	albatross	albatross.n.02
12	almond	almond.n.02
13	alpaca	alpaca.n.03
14	altar	altar.n.01
15	ambulance	ambulance.n.01
16	anchor	anchor.n.01
17	anemometer	anemometer.n.01
18	angelfish	angelfish.n.01
19	anglepoise	floor_lamp.n.01
20	anise	anise.n.02
21	ant	ant.n.01
22	anteater	anteater.n.02
23	antelope	antelope.n.01
24	antenna	antenna.n.01
25	anteroom	anteroom.n.01
26	aphid	aphid.n.01
27	apple	apple.n.01
28	apricot	apricot.n.02
29	apron	apron.n.01
30	aquarium	aquarium.n.01
31	archway	arch.n.03

32	armadillo	armadillo.n.01
33	armchair	armchair.n.01
34	armour	armor.n.01
35	arrow	arrow.n.02
36	artichoke	artichoke.n.02
37	ashtray	ashtray.n.01
38	asparagus	asparagus.n.02
39	aubergine	eggplant.n.01
40	audio-cassette	audiocassette.n.01
41	auger	auger.n.01
42	avocado	avocado.n.01
43	avocet	avocet.n.01
44	awl	awl.n.01
45	ax	ax.n.01
46	axle	axle.n.01
47	babirusa	babirusa.n.01
48	baboon	baboon.n.01
49	backpack	backpack.n.01
50	badge	badge.n.01
51	badger	badger.n.02
52	bagel	bagel.n.01
53	bagpipe	bagpipe.n.01
54	bait	bait.n.02
55	balaclava	balaclava.n.01
56	balalaika	balalaika.n.01
57	balance	balance.n.12
58	ball	ball.n.01
59	ball valve	ball_valve.n.01
60	ballcock	ballcock.n.01
61	baller	scoop.n.06
62	balloon	balloon.n.01
63	balustrade	bannister.n.02
64	banana	banana.n.02

65	bandage	bandage.n.01
66	bandanna	bandanna.n.01
67	bandeau	bandeau.n.01
68	bangle	bangle.n.01
69	banjo	banjo.n.01
70	banquette	banquette.n.01
71	bap	bap.n.01
72	bar	nut_bar.n.01
73	barbell	barbell.n.01
74	barbeque	barbecue.n.03
75	barm	yeast.n.01
76	barn owl	barn_owl.n.01
77	barrel	barrel.n.02
78	baseball	baseball.n.02
79	basin	basin.n.01
80	basket	basket.n.01
81	basketball	basketball.n.01
82	bassoon	bassoon.n.01
83	bat	bat.n.01
84	batfish	batfish.n.01
85	bathtub	bathtub.n.01
86	baton	baton.n.01
87	battery	battery.n.02
88	bayonet	bayonet.n.01
89	beacon	beacon.n.03
90	beaker	beaker.n.01
91	bean	bean.n.01
92	beanbag	beanbag.n.01
93	beans	coffee_bean.n.01
94	bear	bear.n.01
95	beaver	beaver.n.07
96	bed	bed.n.01
97	bedbug	bedbug.n.01

98	bee	bee.n.01
99	beer	beer.n.01
100	beetle	beetle.n.01
101	beetroot	beet.n.02
102	begonia	begonia.n.01
103	bell	bell.n.01
104	bellows	bellows.n.01
105	belt	belt.n.02
106	bench	bench.n.01
107	beret	beret.n.01
108	berry	berry.n.02
109	bike	bicycle.n.01
110	bin	bin.n.01
111	binoculars	binoculars.n.01
112	biplane	biplane.n.01
113	biscuit	biscuit.n.01
114	bison	bison.n.01
115	bit	bit.n.11
116	blackberry	blackberry.n.01
117	blackbird	blackbird.n.02
118	blanket	blanket.n.01
119	blazer	blazer.n.01
120	blender	blender.n.01
121	blenny	blenny.n.01
122	blindfold	blindfold.n.01
123	blinker	blinker.n.02
124	bloodworm	bloodworm.n.01
125	blouse	blouse.n.01
126	blower	blower.n.01
127	bluebell	wild_hyacinth.n.01
128	blueberry	blueberry.n.02
129	bluetit	blue_tit.n.01
130	blunderbuss	blunderbuss.n.01

131	blusher	rouge.n.01
132	boar	boar.n.02
133	bobbin	bobbin.n.01
134	bodkin	bodkin.n.02
135	bogie	tramcar.n.01
136	boiler	boiler.n.01
137	bomb	bomb.n.01
138	bongo	bongo.n.01
139	book	book.n.02
140	bookcase	bookcase.n.01
141	boomerang	boomerang.n.01
142	booth	telephone_booth.n.01
143	boots	boot.n.01
144	bottle	bottle.n.01
145	bouillon	bouillon.n.01
146	bow	bow.n.04
147	bow-tie	bow_tie.n.01
148	bowfin	bowfin.n.01
149	bowl	bowl.n.01
150	box	box.n.01
151	boxfish	boxfish.n.01
152	bra	brassiere.n.01
153	brace	brace.n.07
154	bracelet	bracelet.n.02
155	bracket	bracket.n.04
156	brake	brake.n.01
157	brazier	brazier.n.01
158	bread	bread.n.01
159	brick	brick.n.01
160	bridge	bridge.n.01
161	briefcase	briefcase.n.01
162	brioche	brioche.n.01
163	brocade	brocade.n.01

164	broccoli	broccoli.n.02
165	broom	broom.n.01
166	brush	brush.n.02
167	bucket	bucket.n.01
168	buckle	buckle.n.01
169	budgie	budgerigar.n.01
170	buffalo	old_world_buffalo.n.01
171	bugle	bugle.n.01
172	bulldozer	bulldozer.n.01
173	bullet	bullet.n.01
174	bullion	bullion.n.02
175	bun	bun.n.01
176	bunk	bunk.n.03
177	bus	bus.n.01
178	bustard	bustard.n.01
179	butter	butter.n.01
180	butterfly	butterfly.n.01
181	butterflyfish	flying_gurnard.n.01
182	button	button.n.01
183	butty	butty.n.01
184	buzzard	buzzard.n.02
185	buzzer	buzzer.n.02
186	cab	cab.n.01
187	cabbage	cabbage.n.01
188	cable	cable.n.02
189	cactus	cactus.n.01
190	caecilian	caecilian.n.01
191	cage	cage.n.01
192	cake	cake.n.03
193	calculator	calculator.n.02
194	caliper	caliper.n.01
195	camcorder	camcorder.n.01
196	camel	camel.n.01

197	camera	camera.n.01
198	camper	camper.n.02
199	can-opener	can_opener.n.01
200	canary	canary.n.04
201	candle	candle.n.01
202	candy	candy.n.01
203	canister	canister.n.02
204	cannon	cannon.n.02
205	cannonball	cannonball.n.01
206	canoe	canoe.n.01
207	canopy	canopy.n.03
208	cantaloupe	cantaloup.n.02
209	cap	cap.n.01
210	capacitor	capacitor.n.01
211	capsule	capsule.n.02
212	capybara	capybara.n.01
213	car	car.n.01
214	caramel	caramel.n.02
215	cardigan	cardigan.n.01
216	cardinal	cardinal.n.04
217	carnation	carnation.n.01
218	carousel	carousel.n.02
219	carp	carp.n.02
220	carrot	carrot.n.03
221	cart	cart.n.01
222	cartwheel	cartwheel.n.01
223	cashew	cashew.n.02
224	casket	casket.n.02
225	cassowary	cassowary.n.01
226	castor	caster.n.03
227	cat	cat.n.01
228	catapult	catapult.n.02
229	caterpillar	caterpillar.n.02

230	catfish	catfish.n.03
231	cauldron	caldron.n.01
232	cauliflower	cauliflower.n.02
233	cctv	digital_camera.n.01
234	celeriac	celeriac.n.02
235	celery	celery.n.02
236	cello	cello.n.01
237	cellphone	cellular_telephone.n.01
238	cenotaph	cenotaph.n.01
239	centipede	centipede.n.01
240	cereal	cereal.n.03
241	chainsaw	chain_saw.n.01
242	chair	chair.n.01
243	chaise longue	chaise_longue.n.01
244	chameleon	chameleon.n.03
245	champagne	champagne.n.01
246	chandelier	chandelier.n.01
247	chapati	chapatti.n.01
248	chard	chard.n.02
249	charger	charger.n.02
250	chassis	chassis.n.02
251	cheese	cheese.n.01
252	cheesecake	cheesecake.n.01
253	cheetah	cheetah.n.01
254	cheque	check.n.01
255	cheroot	cheroot.n.01
256	cherry	cherry.n.02
257	chess	chess.n.02
258	chestnut	chestnut.n.03
259	chicken	chicken.n.01
260	chickpea	chickpea.n.03
261	chicory	chicory.n.03
262	chilli	chili.n.02

263	chime	chime.n.01
264	chimney	chimney.n.01
265	chimpanzee	chimpanzee.n.01
266	chinchilla	chinchilla.n.03
267	chinook	chinook.n.05
268	chipmunk	chipmunk.n.01
269	chisel	chisel.n.01
270	chives	chives.n.02
271	chocolate	chocolate.n.02
272	choux	cream_puff.n.01
273	chrysanthemum	chrysanthemum.n.02
274	church	church.n.02
275	cicada	cicada.n.01
276	cichlid	cichlid.n.01
277	cigar	cigar.n.01
278	cigarette	cigarette.n.01
279	cinnamon	cinnamon.n.03
280	clamp	clamp.n.01
281	cleat	cleat.n.02
282	cleaver	cleaver.n.01
283	clipper	clipper.n.04
284	cloche	cloche.n.01
285	clock	clock.n.01
286	clock tower	clock_tower.n.01
287	clog	clog.n.01
288	clove	clove.n.01
289	clutch	clutch.n.05
290	coach	coach.n.04
291	coati	coati.n.01
292	cockatoo	cockatoo.n.01
293	cockchafer	cockchafer.n.01
294	cockroach	cockroach.n.01
295	coconut	coconut.n.02

296	cod	cod.n.03
297	coffee	coffee.n.01
298	coffee-pot	coffeepot.n.01
299	coffin	coffin.n.01
300	coil	coil.n.01
301	coin	coin.n.01
302	colander	colander.n.01
303	coleslaw	coleslaw.n.01
304	collector	solar_array.n.01
305	colobus	colobus.n.01
306	comb	comb.n.01
307	compact-disc	compact_disc.n.01
308	compass	compass.n.01
309	compressor	compressor.n.01
310	concertina	concertina.n.02
311	condom	condom.n.01
312	console	console.n.02
313	container	container.n.01
314	convector	convector.n.01
315	cooker	cooker.n.01
316	cooler	water_cooler.n.01
317	coot	coot.n.01
318	copepod	copepod.n.01
319	copyholder	copyholder.n.01
320	coriander	coriander.n.01
321	corkscrew	corkscrew.n.01
322	cormorant	cormorant.n.01
323	corn	corn.n.03
324	corncrake	corncrake.n.01
325	cornet	cornet.n.01
326	cornflake	corn_flake.n.01
327	cornice	cornice.n.01
328	cos	cos.n.02

329	cot	cot.n.03
330	cotinga	cotinga.n.01
331	cotter	cotter.n.03
332	couch	couch.n.03
333	cougar	cougar.n.01
334	countersink	counterbore.n.01
335	courgette	zucchini.n.02
336	cow	cow.n.02
337	cowbell	cowbell.n.01
338	cowl	hood.n.09
339	cowpea	cowpea.n.01
340	cowrie	cowrie.n.01
341	crab	crab.n.01
342	cracker	cracker.n.01
343	cranberry	cranberry.n.02
344	crank	crank.n.04
345	crate	crate.n.01
346	cricket	cricket_ball.n.01
347	crisp	chip.n.04
348	crispbread	rusk.n.01
349	crocodile	crocodile.n.01
350	croissant	crescent_roll.n.01
351	croquet	croquet.n.01
352	croquette	croquette.n.01
353	crossbill	crossbill.n.01
354	crossbow	crossbow.n.01
355	crouton	crouton.n.01
356	crow	crow.n.01
357	crowbar	crowbar.n.01
358	crown	crown.n.04
359	cuckoo	cuckoo.n.02
360	cucumber	cucumber.n.02
361	cup	cup.n.01

362	cupboard	cupboard.n.01
363	curler	curler.n.01
364	custard	custard.n.01
365	cutter	cutter.n.06
366	cuttlefish	cuttlefish.n.01
367	cylinder	cylinder.n.04
368	cymbal	cymbal.n.01
369	cypress	cypress.n.02
370	daffodil	daffodil.n.01
371	dagger	dagger.n.01
372	dahlia	dahlia.n.01
373	daisy	daisy.n.01
374	damselfish	damselfish.n.01
375	damselfly	damselfly.n.01
376	damson	damson.n.01
377	dart	dart.n.01
378	dartboard	dartboard.n.01
379	date	date.n.08
380	deckchair	deck_chair.n.01
381	deer	deer.n.01
382	deflector	deflector.n.01
383	dentures	denture.n.01
384	depressor	depressor.n.03
385	diamond	diamond.n.02
386	diary	diary.n.02
387	dice	die.n.01
388	dictaphone	dictaphone.n.01
389	dill	dill.n.02
390	dimmer	dimmer.n.01
391	dinosaur	dinosaur.n.01
392	diode	diode.n.02
393	diploma	diploma.n.01
394	dishcloth	dishrag.n.01

395	dishwasher	dishwasher.n.01
396	diskette	diskette.n.01
397	distillery	distillery.n.01
398	divider	divider.n.04
399	dodo	dodo.n.02
400	dog	dog.n.01
401	dogfish	dogfish.n.02
402	doily	doily.n.01
403	dolly	dolly.n.02
404	dolphin	dolphin.n.02
405	dongle	dongle.n.01
406	donkey	domestic_ass.n.01
407	door	door.n.01
408	door-handle	doorknob.n.01
409	doorstop	doorstop.n.01
410	dormouse	dormouse.n.01
411	dough	dough.n.01
412	doughnut	doughnut.n.02
413	dragon	dragon.n.04
414	dragonfly	dragonfly.n.01
415	drain	drain.n.03
416	dresser	dresser.n.05
417	dropper	dropper.n.01
418	drum	drum.n.01
419	duck	duck.n.01
420	dulcimer	dulcimer.n.01
421	dumbbell	dumbbell.n.01
422	durian	durian.n.02
423	duster	duster.n.03
424	eagle	eagle.n.01
425	earphone	earpiece.n.01
426	earplug	earplug.n.01
427	earring	earring.n.01

428	earthworm	earthworm.n.01
429	earwig	earwig.n.01
430	echidna	echidna.n.01
431	eel	eel.n.02
432	egg	egg.n.02
433	eland	eland.n.01
434	elderberry	elderberry.n.02
435	elephant	elephant.n.01
436	elm	elm.n.01
437	emu	emu.n.02
438	engine	engine.n.01
439	envelope	envelope.n.01
440	epee	epee.n.01
441	eraser	eraser.n.01
442	ermine	ermine.n.02
443	escalator	escalator.n.02
444	espadrille	espadrille.n.01
445	extinguisher	fire_extinguisher.n.01
446	fan	fan.n.01
447	faucet	faucet.n.01
448	fax	facsimile.n.02
449	fence	fence.n.01
450	fennel	fennel.n.02
451	fern	fern.n.01
452	ferry	ferry.n.01
453	fez	fez.n.02
454	fife	fife.n.01
455	fig	fig.n.04
456	fighter	fighter.n.02
457	filling-station	gas_station.n.01
458	film	film.n.05
459	finch	finch.n.01
460	fir	fir.n.02

461	firefly	firefly.n.01
462	fireplace	fireplace.n.01
463	fishbowl	fishbowl.n.02
464	fishcake	fish_cake.n.01
465	fishing	fishing.n.01
466	fishing boat	fishing_boat.n.01
467	flambeau	flambeau.n.01
468	flamingo	flamingo.n.01
469	flan	flan.n.01
470	flange	flange.n.01
471	flashlight	flashlight.n.01
472	flathead	flathead.n.02
473	flatworm	flatworm.n.01
474	flea	flea.n.01
475	fleece	fleece.n.03
476	flipper	flipper.n.01
477	flute	flute.n.01
478	fly	fly.n.01
479	flying boat	flying_boat.n.01
480	flywheel	flywheel.n.01
481	foil	foil.n.01
482	folder	booklet.n.01
483	football	soccer_ball.n.01
484	footstool	footstool.n.01
485	forget-me-not	forget-me-not.n.01
486	fork	fork.n.01
487	forklift	forklift.n.01
488	fountain	fountain.n.01
489	fox	fox.n.01
490	foxglove	foxglove.n.01
491	frame	frame.n.10
492	freesia	freesia.n.01
493	freezer	deep-freeze.n.01

494	fridge	fridge.n.01
495	fries	french_fries.n.01
496	frigate	frigate.n.01
497	frock	dress.n.01
498	frog	frog.n.01
499	frying pan	frying_pan.n.01
500	fuse	fuse.n.01
501	galosh	arctic.n.02
502	gannet	gannet.n.01
503	gar	gar.n.01
504	garlic	garlic.n.02
505	gas-mask	gasmask.n.01
506	gate	gate.n.01
507	gavel	gavel.n.01
508	gazebo	gazebo.n.01
509	gazelle	gazelle.n.01
510	gecko	gecko.n.01
511	gerbil	gerbil.n.01
512	gherkin	gherkin.n.02
513	gibbon	gibbon.n.02
514	gift	gift.n.01
515	gimlet	gimlet.n.02
516	ginger	ginger.n.03
517	giraffe	giraffe.n.01
518	glass	glass.n.02
519	glider	glider.n.01
520	glockenspiel	glockenspiel.n.01
521	glove	boxing_glove.n.01
522	glowworm	glowworm.n.01
523	goat	goat.n.01
524	goblet	goblet.n.01
525	goby	goby.n.01
526	goggles	goggles.n.01

527	goldfish	goldfish.n.01
528	golf-ball	golf_ball.n.01
529	golf-club	golf-club.n.01
530	gong	gong.n.01
531	gooseberry	gooseberry.n.02
532	gorilla	gorilla.n.01
533	gourd	gourd.n.02
534	gown	gown.n.05
535	gramophone	gramophone.n.01
536	grape	grape.n.01
537	grapefruit	grapefruit.n.02
538	grapnel	grapnel.n.02
539	grasshopper	grasshopper.n.01
540	grater	grater.n.01
541	graver	graver.n.01
542	gravestone	gravestone.n.01
543	greenfly	greenfly.n.01
544	grenade	grenade.n.01
545	grenadier	grenadier.n.02
546	griddle	griddle.n.01
547	grill	grill.n.02
548	groover	groover.n.01
549	grouper	grouper.n.02
550	grouse	grouse.n.02
551	guacamole	guacamole.n.01
552	guava	guava.n.03
553	guinea fowl	guinea_fowl.n.01
554	guinea pig	guinea_pig.n.02
555	guitar	guitar.n.01
556	gum	chewing_gum.n.01
557	gumboot	rubber_boot.n.01
558	gun	gun.n.01
559	gunnel	gunnel.n.02

560	guppy	guppy.n.01
561	gurnard	gurnard.n.01
562	hacksaw	hacksaw.n.01
563	hadrosaur	hadrosaur.n.01
564	haggis	haggis.n.01
565	hairband	band.n.11
566	hairbrush	hairbrush.n.01
567	hairdryer	hand_blower.n.01
568	halberd	halberd.n.01
569	halibut	halibut.n.02
570	hammer	hammer.n.02
571	hammerhead	hammerhead.n.03
572	hamster	hamster.n.01
573	hand-mirror	hand_glass.n.01
574	handbell	handbell.n.01
575	handcart	handcart.n.01
576	handcuffs	handcuff.n.01
577	handpump	hand_pump.n.01
578	handsaw	handsaw.n.01
579	handset	handset.n.01
580	hanger	hanger.n.02
581	harddisk	hard_disc.n.01
582	harmonica	harmonica.n.01
583	harmonium	harmonium.n.01
584	harp	harp.n.01
585	harpoon	harpoon.n.01
586	harrow	harrow.n.01
587	hasp	hasp.n.01
588	hat	hat.n.01
589	hazelnut	hazelnut.n.01
590	headdress	headdress.n.01
591	headlamp	headlamp.n.01
592	headlight	headlight.n.01

593	headphone	earphone.n.01
594	headset	headset.n.01
595	hearse	hearse.n.01
596	heater	heater.n.01
597	hedgehog	hedgehog.n.02
598	helicopter	helicopter.n.01
599	helmet	helmet.n.01
600	hen	hen.n.01
601	heron	heron.n.02
602	herring	herring.n.02
603	hibiscus	hibiscus.n.01
604	hinge	hinge.n.01
605	hip-flask	hipflask.n.01
606	hippopotamus	hippopotamus.n.01
607	hitch	hitch.n.04
608	hockey	hockey.n.01
609	hoe	hoe.n.01
610	holly	holly.n.01
611	holster	holster.n.01
612	honey	honey.n.01
613	hood	hood.n.08
614	hook	hook.n.04
615	hoop	hoop.n.04
616	hoopo	hoopoe.n.01
617	horn	horn.n.09
618	hornbill	hornbill.n.01
619	horse	horse.n.01
620	horseshoe	horseshoe.n.02
621	hose	hose.n.03
622	hot rod	hot_rod.n.01
623	hotplate	hot_plate.n.01
624	hovercraft	hovercraft.n.01
625	hubcap	hubcap.n.01

626	hummingbird	hummingbird.n.01
627	hummus	hummus.n.01
628	hurdle	hurdle.n.01
629	hut	hut.n.02
630	hyacinth	hyacinth.n.02
631	hyena	hyena.n.01
632	ibex	ibex.n.01
633	ibis	ibis.n.01
634	ice	ice.n.01
635	ice cream	ice_cream.n.01
636	ice-hockey	ice_hockey.n.01
637	ice-pack	ice_pack.n.01
638	iguana	common_iguana.n.01
639	inhaler	inhaler.n.01
640	insect	stick_insect.n.01
641	intercom	intercommunication_system.n.01
642	iris	iris.n.01
643	jack	jack.n.10
644	jackal	jackal.n.01
645	jacket	jacket.n.01
646	jacuzzi	hot_tub.n.01
647	jaguar	jaguar.n.01
648	jam	jam.n.01
649	jar	jar.n.01
650	jay	jay.n.02
651	jeans	jean.n.01
652	jeep	jeep.n.01
653	jellyfish	jellyfish.n.02
654	jet	jet.n.01
655	jig	jig.n.03
656	jug	jug.n.01
657	juice	juice.n.01
658	jumpsuit	jump_suit.n.01

659	junction	junction.n.04
660	kale	kale.n.03
661	kangaroo	kangaroo.n.01
662	kayak	kayak.n.01
663	kazoo	kazoo.n.01
664	kea	kea.n.01
665	keg	keg.n.02
666	kestrel	kestrel.n.02
667	kettle	kettle.n.01
668	key	key.n.01
669	keyboard	keyboard.n.01
670	keyring	key_ring.n.01
671	killifish	killifish.n.01
672	kilt	kilt.n.01
673	kingfisher	kingfisher.n.01
674	kiwi	kiwi.n.03
675	klaxon	klaxon.n.01
676	knife	knife.n.02
677	knob	knob.n.02
678	knuckleduster	brass_knucks.n.01
679	koala	koala.n.01
680	kohlrabi	kohlrabi.n.02
681	kris	kris.n.01
682	kumquat	kumquat.n.01
683	lacewing	lacewing.n.01
684	lacrosse	lacrosse_ball.n.01
685	ladder	ladder.n.01
686	ladle	ladle.n.01
687	ladybird	ladybug.n.01
688	lamp	lamp.n.02
689	lamprey	lamprey.n.01
690	lampshade	lampshade.n.01
691	lancet	lancet.n.02

692	langur	langur.n.01
693	lantern	lantern.n.01
694	lanyard	lanyard.n.02
695	laptop	laptop.n.01
696	lark	lark.n.03
697	latch	latch.n.02
698	lawn mower	lawn_mower.n.01
699	lcd	lcd.n.01
700	leash	leash.n.01
701	leech	leech.n.01
702	leek	leek.n.02
703	lemon	lemon.n.01
704	lemur	lemur.n.01
705	lentil	lentil.n.01
706	leopard	leopard.n.02
707	letterbox	postbox.n.01
708	level	level.n.05
709	licorice	licorice.n.02
710	lifeboat	lifeboat.n.01
711	lift	lift.n.08
712	light	light.n.02
713	light-bulb	light_bulb.n.01
714	lighter	igniter.n.01
715	lilac	lilac.n.01
716	lily	lily.n.01
717	lime	lime.n.06
718	limpet	limpet.n.02
719	lingerie	lingerie.n.01
720	lion	lion.n.01
721	lionfish	lionfish.n.01
722	lipstick	lipstick.n.01
723	lizard	lizard.n.01
724	llama	llama.n.01

725	lobster	lobster.n.02
726	locker	locker.n.01
727	locket	locket.n.01
728	locomotive	locomotive.n.01
729	log	log.n.01
730	lollipop	lollipop.n.02
731	lorikeet	lorikeet.n.01
732	lory	lory.n.01
733	loudspeaker	loudspeaker.n.01
734	lounger	recliner.n.01
735	louse	louse.n.01
736	lumpfish	lumpfish.n.01
737	lupin	lupine.n.01
738	lute	lute.n.02
739	lychee	litchi.n.02
740	lynx	lynx.n.02
741	lyre	lyre.n.01
742	macaque	macaque.n.01
743	macaroni	macaroni.n.02
744	macaroon	macaroon.n.01
745	macaw	macaw.n.01
746	mackerel	mackerel.n.02
747	maggot	maggot.n.01
748	magnet	magnet.n.01
749	magnifier	magnifier.n.01
750	magnolia	magnolia.n.02
751	magpie	magpie.n.01
752	maidenhair	maidenhair.n.01
753	mammoth	mammoth.n.01
754	manatee	manatee.n.01
755	mandolin	mandolin.n.01
756	mango	mango.n.01
757	manta	manta.n.02

758	mantis	mantis.n.01
759	maraca	maraca.n.01
760	margarita	margarita.n.01
761	marigold	marigold.n.01
762	marijuana	marijuana.n.01
763	marmoset	marmoset.n.01
764	marmot	marmot.n.01
765	marrow	marrow.n.02
766	marten	marten.n.01
767	mascara	mascara.n.01
768	masher	masher.n.02
769	mask	mask.n.01
770	mat	mat.n.01
771	matchbox	matchbox.n.01
772	mattress	mattress.n.01
773	matzo	matzo.n.01
774	mayfly	mayfly.n.01
775	mayonnaise	mayonnaise.n.01
776	meatloaf	meat_loaf.n.01
777	medal	decoration.n.02
778	meerkat	meerkat.n.01
779	megaphone	megaphone.n.01
780	meter	meter.n.02
781	metronome	metronome.n.01
782	mic	microphone.n.01
783	microchip	chip.n.07
784	micrometer	micrometer.n.02
785	microscope	microscope.n.01
786	microwave	microwave.n.02
787	midge	midge.n.01
788	millipede	millipede.n.01
789	mink	mink.n.03
790	minnow	minnow.n.01

791	mirror	mirror.n.01
792	missile	missile.n.01
793	mite	mite.n.02
794	mitten	mitten.n.01
795	mixer	mixer.n.04
796	mole	mole.n.06
797	mongoose	mongoose.n.01
798	monkey	monkey.n.01
799	monkfish	goosefish.n.01
800	monoplane	monoplane.n.01
801	moorhen	moorhen.n.01
802	moose	elk.n.01
803	mop	swab.n.02
804	mosquito	mosquito.n.01
805	moth	moth.n.01
806	motherboard	cpu_board.n.01
807	motor	motor.n.01
808	motorboat	motorboat.n.01
809	motorcycle	motorcycle.n.01
810	mouse	mouse.n.01
811	mudskipper	mudskipper.n.01
812	muffin	muffin.n.01
813	mug	mug.n.04
814	mullet	mullet.n.02
815	mushroom	mushroom.n.05
816	mustard	mustard.n.02
817	naan	nan.n.04
818	nail	nail.n.02
819	napkin	napkin.n.01
820	necklace	necklace.n.01
821	needle	needle.n.02
822	nest	nest.n.01
823	nettle	nettle.n.01

824	newspaper	newspaper.n.03
825	nightdress	nightgown.n.01
826	nightingale	nightingale.n.01
827	nipple	nipple.n.02
828	notebook	notebook.n.01
829	nougat	nougat.n.01
830	nutmeg	nutmeg.n.02
831	oar	oar.n.01
832	oat	oat.n.02
833	oboe	oboe.n.01
834	octopus	octopus.n.02
835	odora	alocasia.n.01
836	oil burner	oil_burner.n.01
837	oil lamp	oil_lamp.n.01
838	okapi	okapi.n.01
839	okra	okra.n.01
840	olive	olive.n.04
841	onion	onion.n.03
842	opener	opener.n.03
843	opossum	opossum.n.02
844	orchid	orchid.n.01
845	oriole	old_world_oriole.n.01
846	oscilloscope	oscilloscope.n.01
847	ostrich	ostrich.n.02
848	otter	otter.n.02
849	oven	oven.n.01
850	owl	owl.n.01
851	paddle	paddle.n.04
852	paddlefish	paddlefish.n.01
853	padlock	padlock.n.01
854	pager	beeper.n.01
855	pagoda	pagoda.n.01
856	pail	slop_pail.n.01

857	paintbrush	paintbrush.n.01
858	palette	palette.n.02
859	pancake	pancake.n.01
860	panda	giant_panda.n.01
861	pansy	pansy.n.01
862	panther	panther.n.02
863	panties	pantie.n.01
864	papaya	papaya.n.01
865	paper-bag	shopping_bag.n.01
866	paperclip	paper_clip.n.01
867	paperknife	letter_opener.n.01
868	paprika	paprika.n.02
869	parachute	parachute.n.01
870	parakeet	parakeet.n.01
871	parcel	package.n.02
872	parka	parka.n.01
873	parkia	parkia.n.01
874	parmesan	parmesan.n.01
875	parrot	parrot.n.01
876	parsnip	parsnip.n.03
877	partridge	partridge.n.03
878	pasta	pasta.n.01
879	pastry	pastry.n.02
880	pasty	pasty.n.01
881	pea	pea.n.01
882	peach	peach.n.03
883	peacock	peacock.n.02
884	peanut	peanut.n.01
885	pear	pear.n.02
886	pebble	pebble.n.01
887	pecan	pecan.n.03
888	pedal	pedal.n.02
889	peeler	peeler.n.03

890	peg	peg.n.01
891	pelican	pelican.n.01
892	pellet	pellet.n.02
893	pen	pen.n.01
894	pencil	pencil.n.01
895	penknife	penknife.n.01
896	peony	peony.n.01
897	peppercorn	pepper.n.03
898	peppers	pepper.n.04
899	perch	perch.n.07
900	percolator	percolator.n.01
901	perfume	perfume.n.02
902	persimmon	persimmon.n.02
903	pestle	pestle.n.03
904	pesto	pesto.n.01
905	petunia	petunia.n.01
906	pew	pew.n.01
907	pheasant	pheasant.n.01
908	piano	piano.n.01
909	pick	pick.n.05
910	pick-axe	pick.n.07
911	pickle	pickle.n.01
912	pie	pie.n.01
913	pig	hog.n.03
914	pigeon	pigeon.n.01
915	pike	pike.n.02
916	pillow	pillow.n.01
917	pin	pin.n.09
918	pin-cushion	pincushion.n.01
919	pincer	pincer.n.01
920	pineapple	pineapple.n.02
921	pinto	pinto_bean.n.01
922	pipe	pipe.n.02

923	pipefish	pipefish.n.01
924	piranha	piranha.n.02
925	pistachio	pistachio.n.02
926	pistol	pistol.n.01
927	piston	piston.n.02
928	pitchfork	pitchfork.n.01
929	pitta	flatbread.n.01
930	pizza	pizza.n.01
931	plank	board.n.02
932	plate	plate.n.04
933	platy	platy.n.01
934	pliers	pliers.n.01
935	plover	plover.n.01
936	plow	plow.n.01
937	plug	plug.n.04
938	plum	plum.n.02
939	plunger	plunger.n.03
940	pollack	pollack.n.03
941	pomegranate	pomegranate.n.02
942	pomfret	pomfret.n.01
943	poncho	poncho.n.01
944	popcorn	popcorn.n.02
945	poppy	poppy.n.01
946	porcupine	porcupine.n.01
947	pork	pork.n.01
948	possum	phalanger.n.01
949	pot	pot.n.01
950	potato	potato.n.01
951	potty	chamberpot.n.01
952	pouf	ottoman.n.03
953	pram	baby_buggy.n.01
954	press	press.n.07
955	pressure-cooker	pressure_cooker.n.01

956	pretzel	pretzel.n.01
957	printer	printer.n.03
958	projector	projector.n.02
959	propeller	propeller.n.01
960	pruner	pruner.n.02
961	ptarmigan	ptarmigan.n.01
962	pterodactyl	pterodactyl.n.01
963	puck	puck.n.02
964	pudding	pudding.n.02
965	puffer	puffer.n.02
966	puffin	puffin.n.01
967	pullover	pullover.n.01
968	pump	pump.n.01
969	pumpkin	pumpkin.n.02
970	purse	bag.n.04
971	quad	minibike.n.01
972	quail	quail.n.02
973	quiche	quiche.n.02
974	quill	quill.n.01
975	quilt	quilt.n.01
976	quince	quince.n.02
977	rabbit	rabbit.n.01
978	raccoon	raccoon.n.02
979	racquet	racket.n.04
980	radiator	radiator.n.02
981	radio	radio_receiver.n.01
982	radish	radish.n.01
983	raglan	raglan.n.01
984	railing	railing.n.01
985	raincoat	raincoat.n.01
986	raisin	raisin.n.01
987	rake	rake.n.03
988	rambutan	rambutan.n.02

989	ramp	ramp.n.01
990	rapeseed	rapeseed.n.01
991	rasp	rasp.n.02
992	raspberry	raspberry.n.02
993	rattle	rattle.n.02
994	razor	razor.n.01
995	razorbill	razorbill.n.01
996	reamer	reamer.n.01
997	record player	record_player.n.01
998	recorder	recorder.n.01
999	redcurrant	red_currant.n.01
1000	reeve	reeve.n.01
1001	reflector	reflector.n.01
1002	refrigerator	cooler.n.01
1003	remote control	remote_control.n.01
1004	retractor	retractor.n.01
1005	rhea	rhea.n.02
1006	rhinoceros	rhinoceros.n.01
1007	rhubarb	rhubarb.n.02
1008	rickshaw	jinrikisha.n.01
1009	rifle	rifle.n.01
1010	roach	roach.n.04
1011	roadblock	roadblock.n.02
1012	roaster	roaster.n.04
1013	robe	robe.n.01
1014	robin	robin.n.01
1015	rocker	rocker.n.07
1016	rocket	rocket.n.01
1017	rocking-chair	rocking_chair.n.01
1018	rocking-horse	hobby.n.02
1019	roller	roller.n.01
1020	rolling pin	rolling_pin.n.01
1021	romper	romper.n.02

1022	roof	roof.n.01
1023	rope	rope.n.01
1024	rosary	rosary.n.01
1025	rose	rose.n.01
1026	rosehip	hip.n.05
1027	roulette	roulette.n.02
1028	router	router.n.02
1029	rubber-stamp	stamp.n.09
1030	rudder	rudder.n.01
1031	rugby	rugby_ball.n.01
1032	sachet	sachet.n.01
1033	sack	sack.n.01
1034	saddle	saddle.n.01
1035	safe	safe.n.01
1036	safety pin	safety_pin.n.01
1037	saffron	saffron.n.02
1038	sailboat	sailboat.n.01
1039	sailfish	sailfish.n.02
1040	salad	salad.n.01
1041	salamander	salamander.n.01
1042	salami	salami.n.01
1043	salsa	salsa.n.01
1044	salsify	salsify.n.03
1045	samosa	samosa.n.01
1046	sandal	sandal.n.01
1047	sandpaper	emery_paper.n.01
1048	sandwich	sandwich.n.01
1049	sapphire	sapphire.n.02
1050	satchel	satchel.n.01
1051	satellite	satellite.n.01
1052	saucepan	saucepan.n.01
1053	saucer	saucer.n.02
1054	sauropod	sauropod.n.01

1055	sausage	sausage.n.01
1056	saw	circular_saw.n.01
1057	saxophone	sax.n.02
1058	scale	scale.n.08
1059	scalpel	scalpel.n.01
1060	scanner	scanner.n.02
1061	scarf	scarf.n.01
1062	schoolbus	school_bus.n.01
1063	scissors	scissors.n.01
1064	sconce	sconce.n.04
1065	scone	scone.n.01
1066	scoop	scoop.n.05
1067	scooter	scooter.n.02
1068	scorpion	scorpion.n.03
1069	scraper	scraper.n.01
1070	screamer	screamer.n.03
1071	scroll	scroll.n.02
1072	scythe	scythe.n.01
1073	sea anemone	sea_anemone.n.01
1074	sea urchin	sea_urchin.n.01
1075	seahorse	seahorse.n.02
1076	seal	seal.n.09
1077	sealion	sea_lion.n.01
1078	seaplane	seaplane.n.01
1079	seashell	seashell.n.01
1080	seat	car_seat.n.01
1081	secateurs	secateurs.n.01
1082	seeder	seeder.n.02
1083	sergeant major	sergeant_major.n.02
1084	serow	serow.n.01
1085	set square	set_square.n.01
1086	sewing-machine	sewing_machine.n.01
1087	shaker	shaker.n.03

1088	shark	shark.n.01
1089	sharpener	sharpener.n.01
1090	shaver	shaver.n.03
1091	shears	shears.n.01
1092	sheep	sheep.n.01
1093	shield	shield.n.02
1094	shirt	shirt.n.01
1095	shoe	shoe.n.01
1096	shoelace	shoelace.n.01
1097	shortbread	shortbread.n.01
1098	shorts	short_pants.n.01
1099	shovel	shovel.n.01
1100	shower	shower.n.01
1101	shrew	shrew.n.02
1102	shrimp	shrimp.n.03
1103	shutter	shutter.n.02
1104	shuttlecock	shuttlecock.n.01
1105	sickle	sickle.n.01
1106	side-table	cabinet.n.01
1107	sieve	sieve.n.01
1108	silverfish	silverfish.n.01
1109	simulator	simulator.n.01
1110	sink	sink.n.01
1111	sinker	sinker.n.02
1112	siren	siren.n.04
1113	sitar	sitar.n.01
1114	skate	skate.n.01
1115	skateboard	skateboard.n.01
1116	ski	ski.n.01
1117	skillet	pan.n.01
1118	skimmer	skimmer.n.02
1119	skink	skink.n.01
1120	skirt	skirt.n.02

1121	skullcap	skullcap.n.01
1122	skunk	skunk.n.04
1123	slat	slat.n.01
1124	sled	sled.n.01
1125	sledgehammer	maul.n.01
1126	slicer	slicer.n.03
1127	slide rule	slide_rule.n.01
1128	slingshot	slingshot.n.01
1129	slipper	slipper.n.01
1130	slot	slot.n.07
1131	sloth	sloth.n.02
1132	slug	slug.n.07
1133	snail	snail.n.01
1134	snake	snake.n.01
1135	sneaker	gym_shoe.n.01
1136	snorkel	snorkel.n.01
1137	snowshoe	snowshoe.n.01
1138	snuffer	snuffer.n.01
1139	sock	sock.n.01
1140	socket	socket.n.03
1141	sofa-bed	convertible.n.03
1142	solarium	sun_parlor.n.01
1143	solder	solder.n.01
1144	sombrero	sombrero.n.02
1145	soy	soy.n.04
1146	spade	spade.n.02
1147	spadefoot	spadefoot.n.01
1148	spaghetti	spaghetti.n.01
1149	sparrow	sparrow.n.01
1150	spatula	spatula.n.01
1151	speaker	speaker.n.02
1152	spear	spear.n.01
1153	spectacles	spectacles.n.01

1154	speedometer	speedometer.n.01
1155	sphygmomanometer	sphygmomanometer.n.01
1156	spider	spider.n.01
1157	spinach	spinach.n.02
1158	spinner	spinner.n.02
1159	sponge	sponge.n.01
1160	spoon	spoon.n.01
1161	spoonbill	spoonbill.n.01
1162	sprayer	atomizer.n.01
1163	spring	spring.n.02
1164	sprinkler	sprinkler.n.01
1165	sprocket	sprocket.n.02
1166	sprout	brussels_sprout.n.01
1167	spyglass	field_glass.n.01
1168	squash	squash.n.02
1169	squeegee	squeegee.n.01
1170	squeezer	squeezer.n.01
1171	squid	squid.n.02
1172	squirrel	squirrel.n.01
1173	stamp	stamp.n.04
1174	starfish	starfish.n.01
1175	starling	starling.n.01
1176	steak	steak.n.01
1177	stencil	stencil.n.01
1178	stepper	block.n.01
1179	stereo	stereo.n.01
1180	stethoscope	stethoscope.n.01
1181	stick	stick.n.01
1182	stickleback	stickleback.n.01
1183	stingray	stingray.n.01
1184	stocking	stocking.n.01
1185	stockpot	stockpot.n.01
1186	stonefish	stonefish.n.01

1187	stool	stool.n.01
1188	stopwatch	stopwatch.n.01
1189	stork	stork.n.01
1190	stove	stove.n.02
1191	strainer	strainer.n.01
1192	strawberry	strawberry.n.01
1193	streetlamp	streetlight.n.01
1194	stretcher	stretcher.n.03
1195	sturgeon	sturgeon.n.01
1196	stylus	stylus.n.02
1197	submarine	submarine.n.01
1198	sugar	sugar.n.01
1199	suitcase	bag.n.06
1200	sunfish	sunfish.n.03
1201	sunflower	sunflower.n.01
1202	sunglass	sunglasses.n.01
1203	surfboard	surfboard.n.01
1204	sushi	sushi.n.01
1205	suspenders	brace.n.06
1206	swallow	swallow.n.03
1207	swan	swan.n.01
1208	sweater	sweater.n.01
1209	swimsuit	swimsuit.n.01
1210	swing	swing.n.02
1211	switch	switch.n.01
1212	sword	sword.n.01
1213	sycamore	sycamore.n.03
1214	syringe	syringe.n.01
1215	t-shirt	jersey.n.03
1216	t-square	t-square.n.01
1217	table	table.n.02
1218	table lamp	table_lamp.n.01
1219	tablespoon	tablespoon.n.02

1220	tablet	pill.n.02
1221	tack	tack.n.02
1222	taco	taco.n.02
1223	tadpole	tadpole.n.01
1224	tail-light	taillight.n.01
1225	tamarind	tamarind.n.02
1226	tambourine	tambourine.n.01
1227	tamp	tamp.n.01
1228	tanager	tanager.n.01
1229	tangerine	tangerine.n.01
1230	tank	tank.n.01
1231	tanker	tanker.n.01
1232	tape	tape.n.04
1233	tapir	tapir.n.01
1234	tarantula	tarantula.n.02
1235	taro	taro.n.03
1236	tartlet	tartlet.n.01
1237	tassel	tassel.n.01
1238	taxi	cab.n.03
1239	tea	tea.n.01
1240	teabag	tea_bag.n.01
1241	teapot	teapot.n.01
1242	teaspoon	teaspoon.n.02
1243	teddy	teddy.n.01
1244	telegraph	telegraph.n.01
1245	teleost	teleost_fish.n.01
1246	telephone	telephone.n.01
1247	telescope	telescope.n.01
1248	television	television_receiver.n.01
1249	tench	tench.n.01
1250	tennis	tennis_ball.n.01
1251	tent	tent.n.01
1252	tequila	tequila.n.01

1253	terminal	terminal.n.04
1254	termite	termite.n.01
1255	tern	tern.n.01
1256	tetra	tetra.n.01
1257	thermometer	thermometer.n.01
1258	thermos	thermos.n.01
1259	thermostat	thermostat.n.01
1260	thimble	thimble.n.02
1261	thong	thong.n.02
1262	thread	thread.n.01
1263	throne	throne.n.01
1264	thrush	thrush.n.03
1265	thyme	thyme.n.02
1266	tiara	tiara.n.01
1267	tick	tick.n.02
1268	tie	tie.n.01
1269	tiger	tiger.n.02
1270	timer	timer.n.03
1271	timpani	kettle.n.04
1272	tinamou	tinamou.n.01
1273	toast	toast.n.01
1274	toaster	toaster.n.02
1275	toboggan	toboggan.n.01
1276	toggle	toggle.n.02
1277	toilet	toilet.n.02
1278	token	token.n.03
1279	tomato	tomato.n.01
1280	tongs	tongs.n.01
1281	toothbrush	toothbrush.n.01
1282	toothpick	toothpick.n.01
1283	top-hat	dress_hat.n.01
1284	torch	torch.n.01
1285	torpedo	torpedo.n.03

1286	tortoise	tortoise.n.01
1287	toucan	toucan.n.01
1288	towel	towel.n.01
1289	toy	toy.n.03
1290	tractor	tractor.n.02
1291	trailer	trailer.n.04
1292	trainers	running_shoe.n.01
1293	trampoline	trampoline.n.01
1294	transmitter	transmitter.n.03
1295	tray	tray.n.01
1296	tree	tree.n.01
1297	triangle	triangle.n.05
1298	tribune	tribune.n.02
1299	tricycle	tricycle.n.01
1300	trident	trident.n.01
1301	trimmer	trimmer.n.02
1302	tripod	tripod.n.01
1303	trombone	trombone.n.01
1304	trouser	trouser.n.01
1305	trout	trout.n.02
1306	trowel	trowel.n.01
1307	truck	truck.n.01
1308	truffle	truffle.n.03
1309	trunk	trunk.n.02
1310	tub	tub.n.02
1311	tulip	tulip.n.01
1312	tumbler	tumbler.n.02
1313	tuna	tuna.n.02
1314	turban	turban.n.01
1315	turkey	turkey.n.01
1316	turmeric	turmeric.n.02
1317	turner	turner.n.08
1318	turnip	turnip.n.02

1319	turnstile	turnstile.n.01
1320	turtle	turtle.n.02
1321	tweeter	tweeter.n.01
1322	tweezers	tweezer.n.01
1323	typewriter	typewriter.n.01
1324	ukulele	uke.n.01
1325	umbrella	umbrella.n.01
1326	underpants	underpants.n.01
1327	underwear	underwear.n.01
1328	unicycle	unicycle.n.01
1329	van	van.n.05
1330	vase	vase.n.01
1331	vault	vault.n.02
1332	vervet	vervet.n.01
1333	vest	vest.n.02
1334	video-cassette	videocassette.n.01
1335	video-recorder	videocassette_recorder.n.01
1336	viper	viper.n.01
1337	visor	visor.n.01
1338	volleyball	volleyball.n.02
1339	voltmeter	voltmeter.n.01
1340	vulture	vulture.n.01
1341	waffle	waffle.n.01
1342	wagon	wagon.n.01
1343	waistcoat	vest.n.01
1344	walking-frame	walker.n.05
1345	walkman	walkman.n.01
1346	wallet	wallet.n.01
1347	walnut	walnut.n.01
1348	walrus	walrus.n.01
1349	wand	wand.n.01
1350	warhead	warhead.n.01
1351	warthog	warthog.n.01

1352	washbasin	washbasin.n.01
1353	washing-machine	washing_machine.n.01
1354	wasp	wasp.n.02
1355	wastebin	ashcan.n.01
1356	water-bottle	water_bottle.n.01
1357	watermelon	watermelon.n.02
1358	waveguide	waveguide.n.01
1359	weasel	weasel.n.02
1360	wedding-ring	wedding_ring.n.01
1361	weeder	weeder.n.02
1362	weevil	weevil.n.01
1363	weight	weight.n.02
1364	weka	weka.n.01
1365	well	well.n.01
1366	whale	whale.n.02
1367	wheat	wheat.n.02
1368	wheel	wheel.n.01
1369	wheelbarrow	barrow.n.03
1370	wheelchair	wheelchair.n.01
1371	whip	whip.n.01
1372	whirligig	top.n.08
1373	whisk	whisk.n.01
1374	whistle	whistle.n.03
1375	wigwam	wigwam.n.01
1376	windmill	windmill.n.01
1377	window	window.n.01
1378	wine	wine.n.01
1379	wing-mirror	rearview_mirror.n.01
1380	wiper	windshield_wiper.n.01
1381	wire	wire.n.02
1382	wire cutter	wire_cutter.n.01
1383	wolf	wolf.n.01
1384	wombat	wombat.n.01

1385	woodcock	woodcock.n.01
1386	woodlouse	woodlouse.n.01
1387	woodpecker	woodpecker.n.01
1388	wrasse	wrasse.n.01
1389	wren	wren.n.02
1390	wrench	wrench.n.03
1391	wristband	wristband.n.01
1392	wristwatch	wristwatch.n.01
1393	xylocopa	xylocopa.n.01
1394	xylophone	marimba.n.01
1395	yacht	yacht.n.01
1396	yam	yam.n.03
1397	yarn	yarn.n.02
1398	yo-yo	yo-yo.n.01
1399	yogurt	yogurt.n.01
1400	zebra	zebra.n.01
1401	zip	slide_fastener.n.01
1402	zither	zither.n.01

References

Adlington, R. L., Laws, K. R., & Gale, T. M. (2009). The Hatfield Image Test (HIT): A new picture test and norms for experimental and clinical use. *Journal of clinical and experimental neuropsychology*, 31(6), 731-753.

Allen, I. E., & Seaman, C. A. (2007). Likert scales and data analyses. *Quality progress*, *40*(7), 64-65.

Almuhareb, A. (2006). *Attributes in lexical acquisition* (Doctoral dissertation, University of Essex).

Álvarez, B., & Cuetos, F. (2007). Objective age of acquisition norms for a set of 328 words in Spanish. *Behavior Research Methods*, *39*(3), 377-383.

Andersen, Ø. E., Nioche, J., Briscoe, T., & Carroll, J. A. (2008, May). The BNC parsed with RASP4UIMA. In *LREC*.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of memory and language*, *49*(4), 415-445.

Antony, J. W., & Bennion, K. A. Semantic associates create retroactive interference on an independent spatial recall task. *Experimental Psychology*, *42*(5), 283-290.

Baddeley, A. D., & Dale, H. C. (1966). The effect of semantic similarity on retroactive interference in long-and short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *5*(5), 417-420.

Bakker, A., Krauss, G. L., Albert, M. S., Speck, C. L., Jones, L. R., Stark, C. E., ... & Gallagher, M. (2012). Reduction of hippocampal hyperactivity improves cognition in amnestic mild cognitive impairment. *Neuron*, *74*(3), 467-474.

Barense, M. D., Groen, I. I., Lee, A. C., Yeung, L. K., Brady, S. M., Gregori, M., ... & Henson, R. N. (2012). Intact memory for irrelevant information impairs perception in amnesia. *Neuron*, *75*(1), 157-167.

Barnes, L. L., Schneider, J. A., Boyle, P. A., Bienias, J. L., & Bennett, D. A. (2006). Memory complaints are related to Alzheimer disease pathology in older persons. *Neurology*, *67*(9), 1581-1585.

Baroni, M., Evert, S., & Lenci, A. (2008). Lexical semantics: bridging the gap between semantic theory and computational simulation. In Proceedings of the *ESSLLI Workshop on Distributional Lexical Semantics*.

Baroni, M., Barbu, E., Murphy, B., & Poesio, M. (2010). Strudel: A distributional semantic model based on properties and types. *Cognitive Science*, 34(2): 222-254.

Baroni, M., Dinu, G., & Kruszewski, G. (2014, June). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 238-247).

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(4), 577-660.

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1435), 1177-1187.
Bartko, S. J., Winters, B. D., Cowell, R. A., Saksida, L. M., & Bussey, T. J. (2007). Perirhinal cortex resolves feature ambiguity in configural object recognition and perceptual oddity tasks. *Learning & Memory*, *14*(12), 821-832.
Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48.
Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. *Journal of experimental Psychology*, *80*(3p2), 1.
Bauer, A. J., & Just, M. A. (2017). A brain-based account of "basic-level" concepts. *NeuroImage*, *161*, 196-205.
Baumann, O., Crawshaw, E., & McFadyen, J. (2019). Survival of the Fittest: Increased Stimulus Competition During Encoding Results in Fewer but More Robust Memory Traces. *Frontiers in psychology*, *10*, 21.
Bennett, I. J., Huffman, D. J., & Stark, C. E. (2015). Limbic tract integrity contributes to pattern separation performance across the lifespan. *Cerebral Cortex*, *25*(9), 2988-2999.
Bennett, I. J., & Stark, C. E. (2016). Mnemonic discrimination relates to perforant path integrity: an ultra-high resolution diffusion tensor imaging study. *Neurobiology of learning and memory*, *129*, 107-112.
Bennett, R. W. (1975). Proactive interference in short-term memory: Fundamental forgetting processes. *Journal of Verbal Learning and Verbal Behavior*, *14*(2), 123-144.
Berlin, B. (1972). Speculations on the growth of ethnobotanical nomenclature. *Language in society*, *1*(1), 51-86.
Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 158-167.
Bonnici, H. M., Kumaran, D., Chadwick, M. J., Weiskopf, N., Hassabis, D., & Maguire, E. A. (2012a). Decoding representations of scenes in the medial temporal lobes. *Hippocampus*, *22*(5), 1143-1153.
Bonnici, H. B., Chadwick, M., Kumaran, D., Hassabis, D., Weiskopf, N., & Maguire, E. A. (2012b). Multi-voxel pattern analysis in human hippocampal subfields. *Frontiers in human neuroscience*, *6*, 290.
Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S. Y., Pascual, L., ... & Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child development*, *75*(4), 1115-1139.
Borota, D., Murray, E., Keceli, G., Chang, A., Watabe, J. M., Ly, M., ... & Yassa, M. A. (2014). Post-study caffeine administration enhances memory consolidation in humans. *Nature neuroscience*, *17*(2), 201-203.
Bowman, G. L., Dodge, H. H., Guyonnet, S., Zhou, N., Donohue, J., Bichsel, A., ... & Ousset, P. J. (2019). A blood-based nutritional risk index explains cognitive enhancement and decline in the multidomain

Alzheimer prevention trial. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *5*, 953-963.
Bowman, C. R., & Dennis, N. A. (2015). Age differences in the neural correlates of novelty processing: The effects of item-relatedness. *Brain research*, *1612*, 2-15.
Braak, H., & Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta neuropathologica*, *82*(4), 239-259.
Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, *11*(5), 164-169.
Brainerd, C. J., Yang, Y., Reyna, V. F., Howe, M. L., & Mills, B. A. (2008). Semantic processing in “associative” false memory. *Psychonomic bulletin & review*, *15*(6), 1035-1053.
Brickman, A. M., Khan, U. A., Provenzano, F. A., Yeung, L. K., Suzuki, W., Schroeter, H., ... & Small, S. A. (2014). Enhancing dentate gyrus function with dietary flavanols improves cognition in older adults. *Nature neuroscience*, *17*(12), 1798-1803.
Broadbent, D. E. (1958). Perception and communication. *Pergamon Press, New York.*
Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS one*, *5*(5), e10773.
Brodie, E. E., Wallace, A. M., & Sharrat, B. (1991). Effect of surface characteristics and style of production on naming and verification of pictorial stimuli. *The American journal of psychology*, 517-545.
Bruni, E., Bordignon, U., Liska, A., Uijlings, J., & Sergienya, I. (2013, August). Vsem: An open library for visual semantics representation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 187-192).
Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of artificial intelligence research*, *49*, 1-47.
Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019). English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, *51*(4), 1849-1863.
Buckner, R. L. (2004). Memory and executive function in aging and AD: multiple factors that cause decline and reserve factors that compensate. *Neuron*, *44*(1), 195-208.
Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational linguistics*, *32*(1), 13-47.
Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?
Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior research methods*, *44*(3), 890-907.
Bussey, T. J., & Saksida, L. M. (2002). The organization of visual object representations: a connectionist model of effects of lesions in perirhinal cortex. *European Journal of Neuroscience*, *15*(2), 355-364.
Bussey, T. J., Saksida, L. M., & Murray, E. A. (2002). Perirhinal cortex resolves feature ambiguity in complex visual discriminations. *European Journal of Neuroscience*, *15*(2), 365-374.

Bussey, T. J., Saksida, L. M., & Murray, E. A. (2005). The perceptual-mnemonic/feature conjunction model of perirhinal cortex function. *The Quarterly Journal of Experimental Psychology Section B*, *58*(3-4), 269-282.

Cann, D. R., McRae, K., & Katz, A. N. (2011). False recall in the Deese–Roediger–McDermott paradigm: The roles of gist and associative strength. *Quarterly Journal of Experimental Psychology*, *64*(8), 1515-1542.

Canziani, A., & LeCun, Y. (2020). NYU Deep Learning, Spring 2020.

Capitani, E., Laiacona, M., Mahon, B., & Caramazza, A. (2003). What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. *Cognitive Neuropsychology*, *20*(3-6), 213-261.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in human behavior*, *29*(6), 2156-2160.

Chadwick, M. J., Bonnici, H. M., & Maguire, E. A. (2014). CA3 size predicts the precision of memory recall. *Proceedings of the National Academy of Sciences*, *111*(29), 10720-10725.

Chanales, A. J., Dudukovic, N. M., Richter, F. R., & Kuhl, B. A. (2019). Interference between overlapping memories is predicted by neural states during learning. *Nature communications*, *10*(1), 1-12.

Cheng, N. Y. (1929). Retroactive effect and degree of similarity. *Journal of Experimental Psychology*, *12*(5), 444-449.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior*, *12*(4), 335-359.

Clarke, A., & Tyler, L. K. (2014). Object-specific semantic coding in human perirhinal cortex. *Journal of Neuroscience*, *34*(14), 4766-4775.

Clarke, A., & Tyler, L. K. (2015). Understanding what we see: how we derive meaning from vision. *Trends in cognitive sciences*, *19*(11), 677-687.

Coane, J. H., McBride, D. M., Termonen, M. L., & Cutting, J. C. (2016). Categorical and associative relations increase false memory relative to purely associative relations. *Memory & cognition*, *44*(1), 37-49.

Collell Talleda, G., & Moens, M. F. (2016). Is an image worth more than a thousand words? On the fine-grain semantic differences between visual and linguistic representations. In *Proceedings of the 26th International Conference on Computational Linguistics* (pp. 2807-2817). ACL.

Cowell, R. A., Bussey, T. J., & Saksida, L. M. (2010). Functional dissociations within the ventral object processing pathway: cognitive modules or a hierarchical continuum? *Journal of Cognitive Neuroscience*, *22*(11), 2460-2479.

Cowell, R. A., Bussey, T. J., & Saksida, L. M. (2006). Why does brain damage impair memory? A connectionist model of object recognition memory in perirhinal cortex. *Journal of Neuroscience*, *26*(47), 12186-12197.

Crocco, E., Curiel, R. E., Acevedo, A., Czaja, S. J., & Loewenstein, D. A. (2014). An evaluation of deficits in semantic cueing and proactive and retroactive interference as early features of Alzheimer's disease. *The American Journal of Geriatric Psychiatry*, *22*(9), 889-897.

Cummings, J., Aisen, P., Barton, R., Bork, J., Doody, R., Dwyer, J., ... & Vradenburg, G. (2016). Re-engineering Alzheimer clinical trials: Global Alzheimer's Platform network. *The journal of prevention of Alzheimer's disease*, *3*(2), 114.

Curiel, R. E., Crocco, E., Acevedo, A., Duara, R., Agron, J., & Loewenstein, D. A. (2013). A new scale for the evaluation of proactive and retroactive interference in mild cognitive impairment and early Alzheimer's disease. *Aging*, *1*(1), 1000102.

Curiel, R. E., Crocco, E., Rosado, M., Duara, R., Greig, M. T., Raffo, A., & Loewenstein, D. A. (2016). A brief computerized paired associate test for the detection of mild cognitive impairment in community-dwelling older adults. *Journal of Alzheimer's disease*, *54*(2), 793-799.

Cybenko, A. K. (2011). *Interference in a Modified Recognition Task: An Evaluation of the Changed-trace and Multiple-trace Hypotheses* (Doctoral dissertation, UC Riverside).

Cycowicz, Y. M., Friedman, D., Rothstein, M., & Snodgrass, J. G. (1997). Picture naming by young children: Norms for name agreement, familiarity, and visual complexity. *Journal of experimental child psychology*, *65*(2), 171-237.

Dao, T., & Simpson, T. (2016). WordNet-based semantic similarity measurement. (https://www.codeproject.com/Articles/11835/WordNet-based-semantic-similarity-measurement).

Darby, K., & Sloutsky, V. (2013). Proactive and retroactive interference effects in development. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35, No. 35).

Das, T., Ivleva, E. I., Wagner, A. D., Stark, C. E., & Tamminga, C. A. (2014). Loss of pattern separation performance in schizophrenia suggests dentate gyrus dysfunction. *Schizophrenia research*, *159*(1), 193-197.

Davidoff, J. B., & Ostergaard, A. L. (1988). The role of colour in categorial judgements. *The Quarterly Journal of Experimental Psychology Section A*, *40*(3), 533-544.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255).

Derby, S., Miller, P., Murphy, B., & Devereux, B. (2018). Using sparse semantic embeddings learned from multimodal text and image data to model human conceptual knowledge. *arXiv preprint arXiv:1809.02534*.

Deuker, L., Doeller, C., Fell, J., & Axmacher, N. (2014). Human neuroimaging studies on the hippocampal CA3 region–integrating evidence for pattern separation and completion. *Frontiers in cellular neuroscience*, *8*, 64.

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior research methods*, *46*(4), 1119-1127.

Dewar, M., Pesallaccia, M., Cowan, N., Provinciali, L., & Della Sala, S. (2012). Insights into spared memory capacity in amnestic MCI and Alzheimer's disease via minimal interference. *Brain and cognition*, *78*(3), 189-199.

Donix, M., Ercoli, L. M., Siddarth, P., Brown, J. A., Martin-Harris, L., Burggren, A. C., ... & Bookheimer, S. Y. (2012). Influence of Alzheimer disease family history and genetic risk on cognitive performance in healthy

middle-aged and older people. *The American Journal of Geriatric Psychiatry*, *20*(7), 565-573.
Doraiswamy, P. M., Narayan, V. A., & Manji, H. K. (2018). Mobile and pervasive computing technologies and the future of Alzheimer's clinical trials. *Npj Digital Medicine*, *1*(1), 1-4.
Douglas, D. M., Man, L., Newsome, R. N., Park, H., Aslam, H. M., Barense, M., & Martin, C. B. (2019). Resolving visual and conceptual interference among object concepts requires medial temporal lobe cortex.
Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010, April). Are your participants gaming the system? Screening Mechanical Turk workers. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2399-2402).
Doxey, C. R., & Kirwan, C. B. (2015). Structural and functional correlates of behavioral pattern separation in the hippocampus and medial temporal lobe. *Hippocampus*, *25*(4), 524-533.
Dubuc, M. M., Aubertin-Leheudre, M., & Karelis, A. D. (2020). Relationship between interference control and working memory with academic performance in high school students: The Adolescent Student Academic Performance longitudinal study (ASAP). *Journal of adolescence*, *80*, 204-213.
Dudukovic, N. M., Preston, A. R., Archie, J. J., Glover, G. H., & Wagner, A. D. (2011). High-resolution fMRI reveals match enhancement and attentional modulation in the human medial temporal lobe. *Journal of cognitive neuroscience*, *23*(3), 670-682.
Duff, M. C., Warren, D. E., Gupta, R., Vidal, J. P. B., Tranel, D., & Cohen, N. J. (2012). Teasing apart tangrams: testing hippocampal pattern separation with a collaborative referencing paradigm. *Hippocampus*, *22*(5), 1087-1091.
Duncan, K., Sadanand, A., & Davachi, L. (2012). Memory's penumbra: episodic memory decisions induce lingering mnemonic biases. *Science*, *337*(6093), 485-487.
Ebert, P. L., & Anderson, N. D. (2009). Proactive and retroactive interference in young adults, healthy older adults, and older adults with amnestic mild cognitive impairment. *Journal of the International Neuropsychological Society*, *15*(1), 83-93.
Edland, S. D., Silverman, J. M., Peskind, E. R., Tsuang, D., Wijsman, E., & Morris, J. C. (1996). Increased risk of dementia in mothers of Alzheimer's disease cases: evidence for maternal inheritance. *Neurology*, *47*(1), 254-256.
Erickson, K. I., Leckie, R. L., & Weinstein, A. M. (2014). Physical activity, fitness, and gray matter volume. *Neurobiology of aging*, *35*, S20-S28.
Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E. H., & Smith, N. A. (2015, January). Retrofitting Word Vectors to Semantic Lexicons. In *HLT-NAACL*.
Fellbaum, C. (1998). A semantic network of english: the mother of all WordNets. In *EuroWordNet: A multilingual database with lexical semantic networks* (pp. 137-148). Springer, Dordrecht.
Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001, April). Placing search in context: The concept

revisited. In *Proceedings of the 10th international conference on World Wide Web* (pp. 406-414).
Flynn, D., Van Schaik, P., & Van Wersch, A. (2004). A comparison of multi-item likert and visual analogue scales for the assessment of transactionally defined coping function1. *European Journal of Psychological Assessment*, *20*(1), 49-58.
Fotuhi, M., Hachinski, V., & Whitehouse, P. J. (2009). Changing perspectives regarding late-life dementia. *Nature Reviews Neurology*, *5*(12), 649-658.
Fountain, T., & Lapata, M. (2010). Meaning representation in natural language categorization. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32, No. 32).
French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, *3*(4), 128-135.
Frenck-Mestre, C., & Bueno, S. (1999). Semantic features and semantic categories: Differences in rapid activation of the lexicon. *Brain and language*, *68*(1-2), 199-204.
Frey, A. L., Karran, M., Jimenez, R. C., Baxter, J., Adeogun, M., Bose, N., ... & Routledge, C. (2019). Harnessing the Potential of Digital Technologies for the Early Detection of Neurodegenerative Diseases (edon).
Fujii, T., Saito, D. N., Yanaka, H. T., Kosaka, H., & Okazawa, H. (2014). Depressive mood modulates the anterior lateral CA1 and DG/CA3 during a pattern separation task in cognitively intact individuals: a functional MRI study. *Hippocampus*, *24*(2), 214-224.
Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., ... & Winblad, B. (2006). Mild cognitive impairment. *The lancet*, *367*(9518), 1262-1270.
Gay, E. G. , Weiss, D. J. , Hendel, D. D. , Dawis, R. V. , & Lofquist, L. H. (1971). Manual for the Minnesota importance questionnaire. *Minnesota Studies in Vocational Rehabilitation, 28*, 1-83.
Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic bulletin & review*, *9*(3), 558-565.
Glymour, M. M., Brickman, A. M., Kivimaki, M., Mayeda, E. R., Chêne, G., Dufouil, C., & Manly, J. J. (2018). Will biomarker-based diagnosis of Alzheimer's disease maximize scientific progress? Evaluating proposed diagnostic criteria. *European Journal of Epidemiology*, *33*(7), 607-612.
Grady, C. L., McIntosh, A. R., Horwitz, B., Maisog, J. M., Ungerleider, L. G., Mentis, M. J., ... & Haxby, J. V. (1995). Age-related reductions in human recognition memory due to impaired encoding. *Science*, *269*(5221), 218-221.
Greenleaf, E. A. (1992). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, *29*(2), 176-188.
Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005-10014.
Güsten, J., Ziegler, G., Düzel, E., & Berron, D. (2021). Age impairs mnemonic discrimination of objects more than scenes: A web-based, large-scale approach across the lifespan. *cortex*, *137*, 138-148.

Guyatt, G. H., Townsend, M., Berman, L. B., & Keller, J. L. (1987). A comparison of Likert and visual analogue scales for measuring change in function. *Journal of chronic diseases*, *40*(12), 1129-1133.
Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146-162.
Harvey, P. D., Cosentino, S., Curiel, R., Goldberg, T. E., Kaye, J., Loewenstein, D., ... & Posner, H. (2017). Performance-based and observational assessments in clinical trials across the Alzheimer's disease spectrum. *Innovations in clinical neuroscience*, *14*(1-2), 30.
Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. *Psychology of learning and motivation*, *22*, 193-225.
Haussmann, R., Ganske, S., Gruschwitz, A., Werner, A., Osterrath, A., Lange, J., ... & Donix, M. (2018). Family history of Alzheimer's disease and subjective memory performance. *American Journal of Alzheimer's Disease & Other Dementias®*, *33*(7), 458-462.
Hayden, K. M., Zandi, P. P., West, N. A., Tschanz, J. T., Norton, M. C., Corcoran, C., ... & Cache County Study Group. (2009). Effects of family history and apolipoprotein E ε4 status on cognitive decline in the absence of Alzheimer dementia: The Cache County Study. *Archives of Neurology, 66(11)*, 1378-1383.
Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, *14*(10), e0223792.
Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665-695.
Holden, H. M., Hoebel, C., Loftis, K., & Gilbert, P. E. (2012). Spatial pattern separation in cognitively normal young and older adults. *Hippocampus*, *22*(9), 1826-1832.
Holden, H. M., Toner, C., Pirogovsky, E., Kirwan, C. B., & Gilbert, P. E. (2013). Visual object pattern separation varies in older adults. *Learning & memory*, *20*(7), 358-362.
Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012, July). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 873-882).
Huffman, D. J., & Stark, C. E. (2014). Multivariate pattern analysis of the human medial temporal lobe revealed representationally categorical cortex and representationally agnostic hippocampus. *Hippocampus*, *24*(11), 1394-1403.
Hultsch, D. F., Hertzog, C., & Dixon, R. A. (1984). Text recall in adulthood: The role of intellectual abilities. *Developmental Psychology*, *20*(6), 1193.
Hunsaker, M. R., & Kesner, R. P. (2013). The operation of pattern separation and pattern completion processes associated with different attributes or domains of memory. *Neuroscience & Biobehavioral Reviews*, *37*(1), 36-58.
Jack Jr, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., ... & Trojanowski, J. Q. (2010). Hypothetical model of

dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, *9*(1), 119-128.
Jessen, F., Amariglio, R. E., Van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., ... & Subjective Cognitive Decline Initiative. (2014). A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimer's & dementia*, *10*(6), 844-852.
Jonides, J., Marshuetz, C., Smith, E. E., Reuter-Lorenz, P. A., Koeppe, R. A., & Hartley, A. (2000). Age differences in behavior and PET activation reveal differences in interference resolution in verbal working memory. *Journal of Cognitive Neuroscience*, *12*(1), 188-196.
Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annu. Rev. Psychol.*, *59*, 193-224.
Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in psychology*, *8*, 1726.
Keppel, G., & Underwood, B. J. (1962). Proactive inhibition in short-term retention of single items. *Journal of verbal learning and verbal behavior*, *1*(3), 153-161.
Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology*, *10*(11), e1003915.
Khan, W., Westman, E., Jones, N., Wahlund, L. O., Mecocci, P., Vellas, B., ... & Simmons, A. (2015). Automated hippocampal subfield measures as predictors of conversion from mild cognitive impairment to Alzheimer's disease in two independent cohorts. *Brain topography*, *28*(5), 746-759.
Khanahmadi, M., Farhud, D. D., & Malmir, M. (2015). Genetic of Alzheimer's disease: A narrative review article. *Iranian journal of public health*, *44*(7), 892.
Kiela, D., & Bottou, L. (2014, October). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP)* (pp. 36-45).
Kim, J., & Yassa, M. A. (2013). Assessing recollection and familiarity of similar lures in a behavioral pattern separation task. *Hippocampus*, *23*(4), 287-294.
Kirwan, C. B., & Stark, C. E. (2007). Overcoming interference: An fMRI investigation of pattern separation in the medial temporal lobe. *Learning & Memory*, *14*(9), 625-633.
Kirwan, C. B., Hartshorn, A., Stark, S. M., Goodrich-Hunsaker, N. J., Hopkins, R. O., & Stark, C. E. (2012). Pattern separation deficits following damage to the hippocampus. *Neuropsychologia*, *50*(10), 2408-2414.
Kivisaari, S. L., Tyler, L. K., Monsch, A. U., & Taylor, K. I. (2012). Medial perirhinal cortex disambiguates confusable objects. *Brain*, *135*(12), 3757-3769.
Kivisaari, S. L., Probst, A., & Taylor, K. I. (2013a). The Perirhinal, Entorhinal, and Parahippocampal Cortices and Hippocampus: An Overview of Functional Anatomy and Protocol for Their Segmentation in MR Images. *fMRI*, 239-267.

Kivisaari, S. L., Monsch, A. U., & Taylor, K. I. (2013b). False positives to confusable objects predict medial temporal lobe atrophy. *Hippocampus*, *23*(9), 832-841.
Klementiev, A., Titov, I., & Bhattarai, B. (2012, December). Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012* (pp. 1459-1474).
Koen, J. D., & Rugg, M. D. (2016). Memory reactivation predicts resistance to retroactive interference: evidence from multivariate classification and pattern similarity analyses. *Journal of Neuroscience*, *36*(15), 4389-4399.
Kouststaal, W., Reddy, C., Jackson, E. M., Prince, S., Cendan, D. L., & Schacter, D. L. (2003). False recognition of abstract versus common objects in older and younger adults: testing the semantic categorization account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 499.
Kremer, G., & Baroni, M. (2011). A set of semantic norms for German and Italian. *Behavior Research Methods*, *43*(1), 97-109.
Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*, 417-446.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097-1105.
Krueger, L. E., & Salthouse, T. A. (2010). Differences in acquisition, not retention, largely contribute to sex differences in multitrial word recall performance. *Personality and individual differences*, *49*(7), 768-772.
Krumm, S., Kivisaari, S. L., Probst, A., Monsch, A. U., Reinhardt, J., Ulmer, S., ... & Taylor, K. I. (2016). Cortical thinning of parahippocampal subregions in very early Alzheimer's disease. *Neurobiology of aging*, *38*, 188-196.
Kuhl, B. A., Shah, A. T., DuBrow, S., & Wagner, A. D. (2010). Resistance to forgetting associated with hippocampus-mediated reactivation during new learning. *Nature neuroscience*, *13*(4), 501-506.
Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval research logistics quarterly*, *2*(1-2), 83-97.
Lacy, J. W., Yassa, M. A., Stark, S. M., Muftuler, L. T., & Stark, C. E. (2011). Distinct pattern separation related transfer functions in human CA3/dentate and CA1 revealed using high-resolution fMRI and variable mnemonic similarity. *Learning & memory*, *18*(1), 15-18.
Lancaster, C. L., & Hinds, C. (2019). P1-448: GAMECHANGER: CAN DIGITAL BIOMARKERS TRANSFORM THE DETECTION OF PRECLINICAL ALZHEIMER'S DISEASE?. *Alzheimer's & Dementia*, *15*, P438-P438.
Lancaster, C., Blane, J., Chinner, A., Wolters, L., Koychev, I., & Hinds, C. (2019). The Mezurio smartphone application: Evaluating the feasibility of frequent digital cognitive assessment in the PREVENT dementia study. *MedRxiv*, 19005124.
Lancaster, C., Koychev, I., Blane, J., Chinner, A., Chatham, C., Taylor, K., & Hinds, C. (2020). Gallery Game: Smartphone-based assessment of long-term memory in adults at risk of Alzheimer’s disease. *Journal of clinical and experimental neuropsychology*, *42*(4), 329-343.

Landau, B., Smith, L., & Jones, S. (1998). Object shape, object function, and object name. *Journal of memory and language*, *38*(1), 1-27.
Landauer, T. K. (2007). Lsa as a Theory of Meaning. In *Handbook of Latent Semantic Analysis* (pp. 15-46). Psychology Press.
LaRocque, K. F., Smith, M. E., Carr, V. A., Witthoft, N., Grill-Spector, K., & Wagner, A. D. (2013). Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. *Journal of Neuroscience*, *33*(13), 5466-5474.
Laske, C., Sohrabi, H. R., Frost, S. M., López-de-Ipiña, K., Garrard, P., Buscema, M., ... & O'bryant, S. E. (2015). Innovative diagnostic tools for early detection of Alzheimer's disease. *Alzheimer's & Dementia*, *11*(5), 561-578.
Lazaridou, A., & Baroni, M. (2015). Combining Language and Vision with a Multimodal Skip-gram Model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 153-163).
Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*. The MIT Press, Cambridge, MA, *49*(2), 265-283.
Leal, S. L., & Yassa, M. A. (2014). Effects of aging on mnemonic discrimination of emotional information. *Behavioral neuroscience*, *128*(5), 539.
Leal, S. L., Tighe, S. K., & Yassa, M. A. (2014a). Asymmetric effects of emotion on mnemonic interference. *Neurobiology of learning and memory*, *111*, 41-48.
Leal, S. L., Tighe, S. K., Jones, C. K., & Yassa, M. A. (2014b). Pattern separation of emotional information in hippocampal dentate and CA3. *Hippocampus*, *24*(9), 1146-1155.
Lee, H., Ekanadham, C., & Ng, A. Y. (2007). Sparse deep belief net model for visual area V2. In *Proceedings of the 20th International Conference on Neural Information Processing Systems* (pp. 873-880).
Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.
Lin, D., & Pantel, P. (2001). Discovery of inference rules for question-answering. *Natural Language Engineering*, *7*(4), 343-360.
Liu, K. Y., Gould, R. L., Coulson, M. C., Ward, E. V., & Howard, R. J. (2016). Tests of pattern separation and pattern completion in humans—A systematic review. *Hippocampus*, *26*(6), 705-717.
Loewenstein, D. A., & Acevedo, A. (2005). Training of cognitive and functionally relevant skills in mild Alzheimer’s disease: an integrated approach.
Loewenstein, D. A., Curiel, R. E., Greig, M. T., Bauer, R. M., Rosado, M., Bowers, D., ... & Duara, R. (2016). A novel cognitive stress test for the detection of preclinical Alzheimer disease: discriminative properties and relation to amyloid load. *The American Journal of Geriatric Psychiatry*, *24*(10), 804-813.
Loewenstein, D. A., Acevedo, A., Luis, C., Crum, T., Barker, W. W., & Duara, R. (2004). Semantic interference deficits and the detection of mild Alzheimer's disease and mild cognitive impairment without dementia. *Journal of the International Neuropsychological Society*, *10*(1), 91-100.

Loewenstein, D. A., Acevedo, A., Schram, L., Ownby, R., White, G., Mogosky, B., ... & Duara, R. (2003). Semantic interference in mild Alzheimer disease: preliminary findings. *The American journal of geriatric psychiatry*, *11*(2), 252-255.

Loewenstein, D. A., Curiel, R. E., Duara, R., & Buschke, H. (2018). Novel cognitive paradigms for the detection of memory impairment in preclinical Alzheimer's disease. *Assessment*, *25*(3), 348-359.

Loosli, S. V., Rahm, B., Unterrainer, J. M., Weiller, C., & Kaller, C. P. (2014). Developmental change in proactive interference across the life span: Evidence from two working memory tasks. *Developmental psychology*, *50*(4), 1060.

Loprinzi, P. D., & Frith, E. (2018). The role of sex in memory function: considerations and recommendations in the context of exercise. *Journal of clinical medicine*, *7*(6), 132.

Loprinzi, P. D., Frith, E., Edwards, M. K., Sng, E., & Ashpole, N. (2018). The effects of exercise on memory function among young to middle-aged adults: systematic review and recommendations for future research. *American Journal of Health Promotion*, *32*(3), 691-704.

Lund, F. H. (1926). The criteria of confidence. *The American Journal of Psychology*, *37*(3), 372-381.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological bulletin*, *109*(2), 163.

Maitland, S. B., Herlitz, A., Nyberg, L., Bäckman, L., & Nilsson, L. G. (2004). Selective sex differences in declarative memory. *Memory & cognition*, *32*(7), 1160-1169.

Manelis, A., Paynter, C. A., Wheeler, M. E., & Reder, L. M. (2013). Repetition related changes in activation and functional connectivity in hippocampus predict subsequent memory. *Hippocampus*, *23*(1), 53-65.

Manning, C. D., Raghavan, P., and Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY. 2, 14, 96.

Mark, R. E., & Sitskoorn, M. M. (2013). Are subjective cognitive complaints relevant in preclinical Alzheimer's disease? A review and guidelines for healthcare professionals. *Reviews in Clinical Gerontology*, *23*(1), 61-74.

Martin, E. (1971a). Verbal Learning Theory and Independent Retrieval Phenomena. *Psychological Review*, *78*(4), 314-32.

Martin, E. (1971b). Stimulus component independence. *Journal of Verbal Learning and Verbal Behavior*, *10*(6), 715-721.

Martin, C. B., Sullivan, J. A., Wright, J., & Köhler, S. (2018). How landmark suitability shapes recognition memory signals for objects in the medial temporal lobes. *NeuroImage*, *166*, 425-436.

Martinelli, C., & Shergill, S. S. (2015). Clarifying the role of pattern separation in schizophrenia: the role of recognition and visual discrimination deficits. *Schizophrenia research*, *166*(1-3), 328-333.

McCloy, R., Waugh, G., Medsker, G., Wall, J., Rivkin, D., & Lewis, P. (1999). Development of the O* NET computerized work importance profiler. *Raleigh, NC: National Center for O* NET Development*.

McGeoch, J. A., & McDonald, W. T. (1931). Meaningful relation and retroactive inhibition. *The American Journal of Psychology*, *43*(4), 579-588.

McGeoch, J. A., & Underwood, B. J. (1943). Tests of the two-factor theory of retroactive inhibition. *Journal of Experimental Psychology*, *32*(1), 1.
McGeoch, J. A. (1942). The psychology of human learning.
McGlone, J. (1978). Sex differences in functional brain asymmetry. *Cortex*, *14*(1), 122-128.
McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), 276-282
McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547-559.
McTighe, S. M., Cowell, R. A., Winters, B. D., Bussey, T. J., & Saksida, L. M. (2010). Paradoxical false memory for objects after brain damage. *Science*, *330*(6009), 1408-1410.
Mefoh, P. C. (2010). Gender differences in proactive, retroactive, and no interference conditions. *Gender and Behaviour*, *8*(2), 3036-3047.
Melton, A. W., & Von Lackum, W. J. (1941). Retroactive and proactive inhibition in retention: Evidence for a two-factor theory of retroactive inhibition. *The American Journal of Psychology*, *54*(2), 157-173.
Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 538-558.
Meunier, M., Bachevalier, J., Mishkin, M., & Murray, E. A. (1993). Effects on visual recognition of combined and separate ablations of the entorhinal and perirhinal cortex in rhesus monkeys. *Journal of Neuroscience*, *13*(12), 5418-5432.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes*, *6*(1), 1-28.
Minett, T. S., Dean, J. L., Firbank, M., English, P., & O'Brien, J. T. (2005). Subjective memory complaints, white-matter lesions, depressive symptoms, and cognition in elderly patients. *The American journal of geriatric psychiatry*, *13*(8), 665-671.
Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013). Semantic memory: A feature-based analysis and new norms for Italian. *Behavior research methods*, *45*(2), 440-461.
Montefinese, M., Zannino, G. D., & Ambrosini, E. (2015). Semantic similarity between old and new items produces false alarms in recognition memory. *Psychological research*, *79*(5), 785-794.
Montefinese, M., & Vinson, D. P. (2017). Resemblance among similarity measures in semantic representation. In *CogSci*.
Montefinese, M., Vinson, D., & Ambrosini, E. (2018). Recognition memory and featural similarity between concepts: The pupil's point of view. *Biological psychology*, *135*, 159-169.
Morrow, L. A., Snitz, B. E., Rodriquez, E. G., Huber, K. A., & Saxton, J. A. (2009). High medical co-morbidity and family history of dementia is associated with lower cognitive function in older patients. *Family Practice*, *26*(5), 339-343.
Mortamais, M., Ash, J. A., Harrison, J., Kaye, J., Kramer, J., Randolph, C., ... & Ritchie, K. (2017). Detecting cognitive changes in preclinical Alzheimer's

disease: A review of its feasibility. *Alzheimer's & dementia*, *13*(4), 468-492.
Moss, H. E., Rodd, J. M., Stamatakis, E. A., Bright, P., & Tyler, L. K. (2005). Anteromedial temporal cortex supports fine-grained differentiation among objects. *Cerebral cortex*, *15*(5), 616-627.
Motley, S. E., & Kirwan, C. B. (2012). A parametric investigation of pattern separation processes in the medial temporal lobe. *Journal of Neuroscience*, *32*(38), 13076-13084.
Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, *5*(1), 32-38.
Murdock Jr, B. B. (1964). Proactive inhibition in short-term memory. *Journal of Experimental Psychology*, *68*(2), 184.
Naglieri, J. A., & Rojahn, J. (2001). Gender differences in planning, attention, simultaneous, and successive (PASS) cognitive processes and achievement. *Journal of Educational Psychology*, *93*(2), 430.
Nakov, P., & Hearst, M. A. (2008, June). Solving relational similarity problems using the web as a corpus. In *Proceedings of ACL-08: HLT* (pp. 452-460).
Naveh-Benjamin, M., & Mayr, U. (2018). Age-related differences in associative memory: Empirical evidence and theoretical perspectives. *Psychology and Aging*, *33*(1), 1-6.
Neisser, U. (1987). From direct perception to conceptual structure. *Concepts and conceptual development: Ecological and intellectual factors in categorization*, 11-24.
Nigg, J. T., Blaskey, L. G., Huang-Pollock, C. L., & Rappley, M. D. (2002). Neuropsychological executive functions and DSM-IV ADHD subtypes. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*(1), 59-66.
Nishimoto, T., Miyawaki, K., Ueda, T., Une, Y., & Takahashi, M. (2005). Japanese normative set of 359 pictures. *Behavior research methods*, *37*(3), 398-416.
Nyberg, L., Lövdén, M., Riklund, K., Lindenberger, U., & Bäckman, L. (2012). Memory aging and brain maintenance. *Trends in cognitive sciences*, *16*(5), 292-305.
Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological review*, *56*(3), 132.
Ostergaard, A. L., & Davidoff, J. B. (1985). Some effects of color on naming and recognition of objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(3), 579.
Owens, A. P., Hinds, C., Manyakov, N. V., Stavropoulos, T. G., Lavelle, G., Gove, D., ... & Aarsland, D. (2020). Selecting remote measurement technologies to optimize assessment of function in early Alzheimer's disease: a case study. *Frontiers in psychiatry*, 1163.
Padó, S., Padó, U., & Erk, K. (2007, June). Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 400-409).

Paleja, M., & Spaniol, J. (2013). Spatial pattern completion deficits in older adults. *Frontiers in Aging Neuroscience*, *5*, 3.
Paleja, M., Girard, T. A., Herdman, K. A., & Christensen, B. K. (2014). Two distinct neural networks functionally connected to the human hippocampus during pattern separation tasks. *Brain and cognition*, *92*, 101-111.
Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161-199.
Pavisic, I. M., Nicholas, J. M., Pertzov, Y., O'Connor, A., Liang, Y., Collins, J. D., ... & Crutch, S. J. (2021). Visual short-term memory impairments in presymptomatic familial Alzheimer's disease: A longitudinal observational study. *Neuropsychologia*, *162*, 108028.
Perfetti, C. A. (1998). The Limits of Co-Occurrence: Tools and Theories in Language Research. *Discourse Processes*, *25*, 363-377.
Perrotin, A., Mormino, E. C., Madison, C. M., Hayenga, A. O., & Jagust, W. J. (2012). Subjective cognition and amyloid deposition imaging: a Pittsburgh Compound B positron emission tomography study in normal elderly individuals. *Archives of neurology*, *69*(2), 223-229.
Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of neurology*, *56*(3), 303-308.
Picker, C. J. (2015). Retroactive interference in recognition memory: The effects of mental effort and similarity on recollection and familiarity.
Pidgeon, L. M., & Morcom, A. M. (2014). Age-related increases in false recognition: the role of perceptual and conceptual similarity. *Frontiers in Aging Neuroscience*, *6*, 283.
Purser, J. L., Fillenbaum, G. G., & Wallace, R. B. (2006). Memory complaint is not necessary for diagnosis of mild cognitive impairment and does not predict 10-year trajectories of functional disability, word recall, or short portable mental status questionnaire limitations. *Journal of the American Geriatrics Society*, *54*(2), 335-338.
Qian, J., Wolters, F. J., Beiser, A., Haan, M., Ikram, M. A., Karlawish, J., ... & Blacker, D. (2017). APOE-related risk of mild cognitive impairment and dementia for prevention trials: an analysis of four cohorts. *PLoS medicine*, *14*(3), e1002254.
R Core Team. (2021, version 1.1.25). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/
Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics*, *19*(1), 17-30.
Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*(33), 7255-7269.
Reagh, Z. M., & Yassa, M. A. (2014). Object and spatial mnemonic interference differentially engage lateral and medial entorhinal cortex in humans. *Proceedings of the National Academy of Sciences*, *111*(40), E4264-E4273.

Reagh, Z. M., Roberts, J. M., Ly, M., DiProspero, N., Murray, E., & Yassa, M. A. (2014). Spatial discrimination deficits as a function of mnemonic interference in aged adults with and without memory impairment. *Hippocampus*, *24*(3), 303-314.

Rei, M., & Briscoe, T. (2014, June). Looking for hyponyms in vector space. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning* (pp. 68-77).

Reid, L. M., & MacLullich, A. M. (2006). Subjective memory complaints and cognitive impairment in older people. *Dementia and geriatric cognitive disorders*, *22*(5-6), 471-485.

Reisinger, J., & Mooney, R. (2010). A mixture model with sharing for lexical semantics. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 1173-1182).

Resnik, P. (1995, August). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 1* (pp. 448-453).

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303-345.

Roberts, J. M., Ly, M., Murray, E., & Yassa, M. A. (2014). Temporal discrimination deficits as a function of lag interference in older adults. *Hippocampus*, *24*(10), 1189-1196.

Robinson, E.S. (1920). Studies from the psychological laboratory of the University of Chicago: Some factors determining the degree of retroactive inhibition. *Psychological Monographs*, 28, 1-57.

Robinson, E. S. (1927). The'similarity' factor in retroaction. *The American Journal of Psychology*, *39*, 297-312.

Rogers, T. T., & Plaut, D. C. (2002). Connectionist perspectives on category-specific deficits. *Category-specificity in brain and mind*, 251-290.

Roman, S., Axler, S., & Gehring, F. W. (2005). *Advanced linear algebra* (Vol. 3). New York: Springer.

Rosch, E., & Lloyd, B. B. (Eds.). (1978). Cognition and categorization.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382-439.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*(2), 217-236.

Rowe, C. C., Ellis, K. A., Rimajova, M., Bourgeat, P., Pike, K. E., Jones, G., ... & Villemagne, V. L. (2010). Amyloid imaging results from the Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging. *Neurobiology of aging*, *31*(8), 1275-1283.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627-633.

Rucklidge, J. J., & Tannock, R. (2002). Neuropsychological profiles of adolescents with ADHD: Effects of reading difficulties and gender. *Journal of child psychology and psychiatry*, *43*(8), 988-1003.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211-252.
Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers*, 36(3), 506-515.
Saksida, L. M., & Bussey, T. J. (2010). The representational–hierarchical view of amnesia: Translation from animal to human. *Neuropsychologia*, *48*(8), 2370-2384.
Saltz, E., & Hamilton, H. (1967). Spontaneous recovery of List 1 responses in the AB, A'-C paradigm. *Journal of Experimental Psychology*, *75*(2), 267.
Samrani, G., & Persson, J. (2021). Proactive interference in working memory is related to adult age and cognitive factors: cross-sectional and longitudinal evidence from the Betula study. *Aging, Neuropsychology, and Cognition*, *28*(1), 108-127.
Samrani, G., Bäckman, L., & Persson, J. (2017). Age-Differences in the Temporal Properties of Proactive Interference in Working Memory. *Psychology and Aging*, *32*(8), 722-731.
Schacter, D. L., Cooper, L. A., & Valdiserri, M. (1992). Implicit and explicit memory for novel visual objects in older and younger adults. *Psychology and aging*, *7*(2), 299.
Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica neerlandica*, *57*(1), 19-35.
Schmeidler, G. R. (1939). Retroaction and proaction in serial learning. *The American Journal of Psychology*, *52*(4), 592-600.
Schmiedek, F., Li, S. C., & Lindenberger, U. (2009). Interference and facilitation in spatial working memory: age-associated differences in lure effects in the n-back paradigm. *Psychology and aging*, *24*(1), 203.
Segal, S. K., Stark, S. M., Kattan, D., Stark, C. E., & Yassa, M. A. (2012). Norepinephrine-mediated emotional arousal facilitates subsequent pattern separation. *Neurobiology of learning and memory*, *97*(4), 465-469.
Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, *29*(3), 411-426.
Shelton, D. J., & Kirwan, C. B. (2013). A possible negative influence of depression on the ability to overcome memory interference. *Behavioural brain research*, *256*, 20-26.
Sheppard, D. P., Graves, L. V., Holden, H. M., Delano-Wood, L., Bondi, M. W., & Gilbert, P. E. (2016). Spatial pattern separation differences in older adult carriers and non-carriers for the apolipoprotein E epsilon 4 allele. *Neurobiology of learning and memory*, *129*, 113-119.
Shimamura, A. (1994). Memory and frontal lobe function. The Cognitive Neurosciences, 803-813.
Silberer, C., & Lapata, M. (2014, June). Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 721-732).

Silberer, C. H. (2015). *Learning visually grounded meaning representations* (Doctoral dissertation, University of Edinburgh).
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
Skaggs, E.B. (1925). Further studies in retroactive inhibition. Psychological Monographs, 39, 25-32.
Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory*, *6*(2), 174.
Sosic-Vasic, Z., Hille, K., Kröner, J., Spitzer, M., & Kornmeier, J. (2018). When learning disturbs memory–Temporal profile of retroactive interference of learning on memory formation. *Frontiers in psychology*, *9*, 82.
South, M., Stephenson, K. G., Nielson, C. A., Maisel, M., Top, D. N., & Kirwan, C. B. (2015). Overactive pattern separation memory associated with negative emotionality in adults diagnosed with autism spectrum disorder. *Journal of autism and developmental disorders*, *45*(11), 3458-3467.
Sperling, R. A., Dickerson, B. C., Pihlajamaki, M., Vannini, P., LaViolette, P. S., Vitolo, O. V., ... & Johnson, K. A. (2010). Functional alterations in memory networks in early Alzheimer's disease. *Neuromolecular medicine*, *12*(1), 27-43.
Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., ... & Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, *7*(3), 280-292.
Sperling, R. A., Rentz, D. M., Johnson, K. A., Karlawish, J., Donohue, M., Salmon, D. P., & Aisen, P. (2014). The A4 study: stopping AD before symptoms begin? *Science translational medicine*, *6*(228), 228fs13-228fs13.
Spielberg, J. M., Galarce, E. M., Ladouceur, C. D., McMakin, D. L., Olino, T. M., Forbes, E. E., ... & Dahl, R. E. (2015). Adolescent development of inhibition as a function of SES and gender: Converging evidence from behavior and fMRI. *Human brain mapping*, *36*(8), 3194-3203.
Staresina, B. P., Cooper, E., & Henson, R. N. (2013). Reversible information flow across the medial temporal lobe: the hippocampus links cortical modules during memory retrieval. *Journal of Neuroscience*, *33*(35), 14184-14192.
Stark, S. M., Yassa, M. A., Lacy, J. W., & Stark, C. E. (2013). A task to assess behavioral pattern separation (BPS) in humans: Data from healthy aging and mild cognitive impairment. *Neuropsychologia*, *51*(12), 2442-2449.
Stark, S. M., Stevenson, R., Wu, C., Rutledge, S., & Stark, C. E. (2015). Stability of age-related deficits in the mnemonic similarity task across task variations. *Behavioral neuroscience*, *129*(3), 257.
Stark, S. M., Kirwan, C. B., & Stark, C. E. (2019). Mnemonic similarity task: A tool for assessing hippocampal integrity. *Trends in cognitive sciences*, *23*(11), 938-951.

Sung, Y. T., & Wu, J. S. (2018). The visual analogue scale for rating, ranking and paired-comparison (VAS-RRP): a new technique for psychological measurement. *Behavior research methods*, *50*(4), 1694-1715.
Sussna, M. J. (1997). *Text retrieval using inference in semantic metanetworks*. University of California, San Diego.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).
Szumlanski, S., Gomez, F., & Sims, V. K. (2013, August). A new set of norms for semantic relatedness measures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 890-895).
Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology*, *23*(3), 457-482.
Taylor, K. I., Stamatakis, E. A., & Tyler, L. K. (2009). Crossmodal integration of object features: voxel-based correlations in brain-damaged patients. *Brain*, *132*(3), 671-683.
Taylor, K. I., Devereux, B. J., Acres, K., Randall, B., & Tyler, L. K. (2012). Contrasting effects of feature-based statistics on the categorisation and basic-level identification of visual objects. *Cognition*, *122*(3), 363-374.
Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184-197.
Tolentino, J. C., Pirogovsky, E., Luu, T., Toner, C. K., & Gilbert, P. E. (2012). The effect of interference on temporal order memory for random and fixed sequences in nondemented older adults. *Learning & Memory*, *19*(6), 251-255.
Toner, C. K., Pirogovsky, E., Kirwan, C. B., & Gilbert, P. E. (2009). Visual object pattern separation deficits in nondemented older adults. *Learning & memory*, *16*(5), 338-342.
Toner, C. K., Reese, B. E., Neargarder, S., Riedel, T. M., Gilmore, G. C., & Cronin-Golomb, A. (2012). Vision-fair neuropsychological assessment in normal aging, Parkinson's disease and Alzheimer's disease. *Psychology and aging*, *27*(3), 785.
Turian, J., Ratinov, L., & Bengio, Y. (2010, July). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 384-394).
Turner RC. (2006). Alzheimer's disease. *Seminars in Neurol*, 26: 499–506.
Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*, 141-188.
Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, *32*(3), 379-416.
Tyler, L. K., Stamatakis, E. A., Bright, P., Acres, K., Abdallah, S., Rodd, J. M., & Moss, H. E. (2004). Processing objects at different levels of specificity. *Journal of cognitive neuroscience*, *16*(3), 351-362.

Tyler, L. K., Chiu, S., Zhuang, J., Randall, B., Devereux, B. J., Wright, P., ... & Taylor, K. I. (2013). Objects and categories: feature statistics and object processing in the ventral stream. *Journal of cognitive neuroscience*, *25*(10), 1723-1735.
Underwood, B. J. (1945). The effect of successive interpolations on retroactive and proactive inhibition. *Psychological Monographs*, *59*(3), i.
Ungerleider, L. G., Mishkin, M. (1982). Two cortical visual systems. In: Ingle, D. J., Goodale, M., Mansfield, R. J. W. (eds.). *Analysis of Visual Behaviour*. Cambridge, MA: MIT Press, 549-586.
van Kesteren, M. T., Brown, T. I., & Wagner, A. D. (2018). Learned Spatial Schemas and Prospective Hippocampal Activity Support Navigation After One-Shot Learning. *Frontiers in human neuroscience*, *12*, 486.
van Mourik, R., Oosterlaan, J., & Sergeant, J. A. (2005). The Stroop revisited: A meta-analysis of interference control in AD/HD. *Journal of Child Psychology and Psychiatry*, *46*(2), 150-165.
Vieweg, P., Stangl, M., Howard, L. R., & Wolbers, T. (2015). Changes in pattern completion–a key mechanism to explain age-related recognition memory deficits? *Cortex*, *64*, 343-351.
Viggiano, M. P., Vannucci, M., & Righi, S. (2004). A new standardized set of ecological pictures for experimental and clinical research on visual object processing. *Cortex*, *40*(3), 491-509.
Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive psychology*, *48*(4), 422-488.
Vinson, D. P., & Vigliocco, G. (2008). Feature norms for a large set of object and event concepts. *Behavior Research Methods*, 40, 183-190.
Vivas, J., Vivas, L., Comesaña, A., Coni, A. G., & Vorano, A. (2017). Spanish semantic feature production norms for 400 concrete concepts. *Behavior Research Methods*, 49(3), 1095-1106.
von Gunten, A., & Ron, M. A. (2004). Hippocampal volume and subjective memory impairment in depressed patients. *European psychiatry*, *19*(7), 438-440.
Wild, K. V., Mattek, N., Austin, D., & Kaye, J. A. (2016). “Are you sure?” Lapses in self-reported activities among healthy older adults reporting online. *Journal of Applied Gerontology*, *35*(6), 627-641.
Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*, 235-269.
Wixted, J. T. (2005). A theory about why we forget what we once knew. *Current Directions in Psychological Science*, *14*(1), 6-9.
Wolf, M. S., Gazmararian, J. A., & Baker, D. W. (2005). Health literacy and functional health status among older adults. *Archives of internal medicine*, *165*(17), 1946-1952.
Woud, M. L., Heeren, A., Shkreli, L., Meyer, T., Egeri, L., Cwik, J. C., ... & Margraf, J. (2019). Investigating the effect of proactive interference control training on intrusive memories. *European journal of psychotraumatology*, *10*(1), 1611092.
Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.
Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural

responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619-8624.
Yassa, M. A., Mattfeld, A. T., Stark, S. M., & Stark, C. E. (2011a). Age-related memory deficits linked to circuit-specific disruptions in the hippocampus. *Proceedings of the National Academy of Sciences*, *108*(21), 8873-8878.
Yassa, M. A., Lacy, J. W., Stark, S. M., Albert, M. S., Gallagher, M., & Stark, C. E. (2011b). Pattern separation deficits associated with increased hippocampal CA3 and dentate gyrus activity in nondemented older adults. *Hippocampus*, *21*(9), 968-979.
Yong, C. & Foo., S.K. (1999). A case study on inter-annotator agreement for word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Standardizing Lexical Resources (SIGLEX99)*.
Zeiler, M. D., & Fergus, R. (2013). Stochastic pooling for regularization of deep convolutional neural networks. *arXiv preprint arXiv:1301.3557*.
Zelinski, E. M., & Gilewski, M. J. (1988). Memory for prose and aging: A meta-analysis. In *Cognitive development in adulthood* (pp. 133-158). Springer, New York, NY.
Zelinski, E. M., Burnight, K. P., & Lane, C. J. (2001). The relationship between subjective and objective memory in the oldest old: Comparisons of findings from a representative and a convenience sample. *Journal of Aging and Health*, *13*(2), 248-266.
Zola-Morgan, S., Squire, L. R., Amaral, D. G., & Suzuki, W. A. (1989). Lesions of perirhinal and parahippocampal cortex that spare the amygdala and hippocampal formation produce severe memory impairment. *Journal of Neuroscience, 9(12)*, 4355-4370.