Interactions Between Memories During Learning: A Behavioural Measure of Representational Plasticity

by

Youssef Ali

A thesis submitted to the Graduate Program in Psychology
in conformity with the requirements for
the Degree of Master of Science

Queen’s University
Kingston, Ontario, Canada
(September, 2022)

Copyright ©Youssef Ali, 2022
Abstract

The capacity of human memory is enormous, but not unlimited. Because of the finite neural tissue in the hippocampus, memories are often encoded with overlapping patterns of activation, which can facilitate the cohesion of related memories, but also evoke competition between them. One theory of neuroplasticity, the non-monotonic plasticity hypothesis (NMPH), predicts that the initial degree of overlap or co-activation amongst memories may drive learning-related change; memories integrate at high co-activation, but due to competition, differentiate at moderate co-activation. As an alternate theory, recent evidence suggests a competitive differentiation (CD) mechanism whereby greater memory co-activation results in differentiation, not integration. We bridge these hypotheses by curating a stimulus set that samples a comprehensive range of visual similarity as a proxy for memory co-activation, and by using participants’ selections of incorrect memory lures as a novel behavioural index of integration and differentiation. Because they are theorized to differentially recruit trisynaptic and monosynaptic hippocampal pathways, respectively, we measured patterns of representational change in two major learning paradigms: episodic and statistical learning. We used a convolutional neural network to generate image pairs, which were embedded into a sequence where participants learned paired associates either episodically or statistically. Representational change was measured using a four-alternative forced choice task. Importantly, none of the options were the correct pairmate; rather, they were subtly shifted to be either more or less visually similar to the cue, signaling integration or differentiation respectively. Statistical learning followed an NMPH-consistent pattern, while episodic learning followed a more idiosyncratic pattern, where integration was attenuated at higher co-activation. When the experiment was repeated with a narrower similarity range, statistical learning followed a linear pattern that fits into the broader NMPH pattern, whereas
episodic learning again showed integration attenuation at the end of the range, inconsistent with the broader pattern. This may suggest an additive influence of NMPH and CD at high co-activation. We also found that differentiation occurs at a lower level of co-activation in episodic compared to statistical learning. This work established a novel behavioural task for measuring subtle shifts in memory, and builds toward a general framework for learning-related representational change.
Co-Authorship

This thesis was written in collaboration with Jeffrey Wammes, PhD, Assistant Professor, Department of Psychology, Queen’s University.
Acknowledgements

The following work would not have been possible without the disciplined instruction of my supervisor, collaborator, and friend, Dr. Jeffrey Wammes, whose mentorship and authenticity have proven indispensable to this project. I also extend my thanks to my examiners, Dr. Jason Gallivan, Dr. Hans Dringenberg, and Dr. Gavin Winston for volunteering their valuable time and feedback for this thesis. Most importantly, I offer my sincerest gratitude to my parents, Fatma and Ayman, whose moral excellence humbles me, and whose support and encouragement hearten me.
# Table of Contents

Abstract ......................................................................................................................................... ii
Co-Authorship ................................................................................................................................. iv
Acknowledgements ......................................................................................................................... v
Table of Contents ........................................................................................................................... vi
List of Tables .................................................................................................................................. ix
List of Figures ................................................................................................................................. x
List of Abbreviations and Symbols ................................................................................................. xi
Introduction ...................................................................................................................................... 1

- Synaptic plasticity ......................................................................................................................... 1
- From single neurons to populations ......................................................................................... 2
- Representational Change ............................................................................................................ 4
- Competitive Differentiation ....................................................................................................... 7
- Reconciliation .............................................................................................................................. 8
- The Present Work ....................................................................................................................... 10

Experiment 1 .................................................................................................................................. 11
Methods .......................................................................................................................................... 12

- Participants ................................................................................................................................. 12
- Materials .................................................................................................................................... 12
- Procedure .................................................................................................................................. 15
  - Learning Task ........................................................................................................................ 15
  - Four-alternative Forced Choice Task .................................................................................... 17
- Analysis Approach ...................................................................................................................... 19
- Exclusions .................................................................................................................................. 20

Results and Discussion .................................................................................................................. 21
Episodic Learning ............................................................................................................................. 21

- Mixed Effects Model .................................................................................................................. 21
- Bootstrapping Model .................................................................................................................. 24
- Contrasts .................................................................................................................................... 24

Statistical Learning ......................................................................................................................... 26

- Mixed Effects Model .................................................................................................................. 26
- Bootstrapping Model .................................................................................................................. 28
- Contrasts .................................................................................................................................... 28
List of Tables

Table 1 Experiment 1 Episodic Condition Fit Quality ................................................. 22
Table 2 Experiment 1 Statistical Condition Fit Quality .................................................. 27
Table 3 Experiment 2 Episodic Condition Fit Quality ................................................... 39
Table 4 Experiment 3 Episodic Condition Fit Quality ................................................... 46
Table 5 Experiment 3 Statistical Condition Fit Quality .................................................. 50
List of Figures

Figure 1 Feature-Dependent Co-activation ................................................................. 3
Figure 2 NMPH versus CD .......................................................................................... 5
Figure 3 Image Synthesis Procedure ........................................................................... 15
Figure 4 Learning and Outcome Tasks ........................................................................ 17
Figure 5 Experiment 1 Representational Change in Episodic and Statistical Learning .... 23
Figure 6 Experiment 1 Within-Condition Contrasts of Representational Shift ............. 26
Figure 7 Experiment 2 Similarity Level Accuracy Compared to Chance ....................... 37
Figure 8 Experiment 2 Representational Change in Episodic Learning ....................... 40
Figure 9 Experiment 2 Episodic Contrasts of Representational Shift ........................... 42
Figure 10 Experiment 3 Representational Change in Episodic and Statistical Learning .... 47
Figure 11 Experiment 3 Within-Condition Contrasts of Representational Shift ............. 49
Figure 12 Experiments 1 & 3 Representational Change in Episodic and Statistical Learning ... 52
Figure 13 Hybrid Model of Representational Change as a Function of Co-activation ........ 59
List of Abbreviations and Symbols

µ ........................................................................................................... Half-Gaussian Distribution Peak
4AFC ................................................................................................. Four-Alternative Forced Choice
BCM ................................................................................................. Bienenstock-Cooper-Munro
CA ........................................................................................................ Cornu Ammonis
CD ......................................................................................................... Competitive Differentiation
CI .......................................................................................................... Confidence Interval
CNN ..................................................................................................... Convolutional Neural Network
DCT ....................................................................................................... Distributed Coding Theory
df .......................................................................................................... Degrees of Freedom
DG .......................................................................................................... Dentate Gyrus
fMRI .................................................................................................... Functional Magnetic Resonance Imaging
LTD ....................................................................................................... Long-Term Depression
LTP ....................................................................................................... Long-Term Potentiation
M ........................................................................................................... Mean
MSP ....................................................................................................... Monosynaptic Pathway
N ........................................................................................................... Sample Size
NMPH ............................................................................................... Non-monotonic Plasticity Hypothesis
p ........................................................................................................... Probability Value
SD ......................................................................................................... Standard Deviation
TSP ....................................................................................................... Trisynaptic Pathway
wAIC ................................................................................................. Weighted Akaike Information Criterion
wBIC ................................................................................................. Weighted Bayesian Information Criterion
σ ........................................................................................................... Half-Gaussian Distribution Spread
χ² .......................................................................................................... Chi-Squared Statistic
Introduction

Effectively organizing our memories is vital to our interactions with and interpretations of the world. If we cannot learn from past events, we would not be able to recall the faces of friends, where we parked our cars, or the dangers of touching an open fire. The hippocampus, one of the brain’s oldest and most well-studied structures, has been implicated primarily in learning and memory. However, its mechanisms for encoding the relationships between associated memories remain elusive. How strongly must multiple memories be related to one another to comprise a single holistic memory? How do memories move closer or further from each other in an abstract representational space? Can the recollection of one memory attenuate or enhance the recollection of another? And most importantly, can the task context in which we retrieve memories influence the extent to which they overlap with one another? In this study, we use model-based stimulus generation and novel behavioural tasks to address these questions and gain a better understanding of the dynamics of relational memories encoded either statistically or episodically. Specifically, we aim to examine the fate of associated and overlapping memories following learning, to determine whether their underlying memory representations change to become either more, or less, similar to one another.

Synaptic plasticity

While the present thesis is concerned with measuring a relatively coarse behavioural readout of changes in representations, the foundations of representational dynamics in memory can be traced back to the much more fine-grained level of the synapse. Specifically, individual synaptic connections are either strengthened or weakened based on the degree of coincident pre- and postsynaptic activation: long-term potentiation (LTP) refers to a strengthening of a synaptic connection as a result of high-frequency stimulation at a synapse, while long-term depression
(LTD) refers to a weakening of a connection as a result of low- or moderate-frequency stimulation (Bliss & Lomo 1973, Hebb, 1949). This indicates that the frequency of stimulation predicts synaptic strength in a U-shaped function, where low excitation yields no change, moderate yields weakening, and high yields strengthening of the synapse. According to Bienenstock-Cooper-Munro (BCM) theory (Bienenstock et al., 1982), the level of excitation at which a neuron switches from weakening (i.e., LTD) to strengthening (i.e., LTP) is a dynamically changing threshold, which appears to be heavily experience-dependent. This means that the changes in synaptic strength could vary as a function of task, stimulus or context. For instance, when rats are light-deprived, they show greater sensitivity in the visual cortex which in turn can be associated with an increased likelihood of synaptic strengthening and a decreased likelihood of weakening (Kirkwood et al., 1996). It is clear from these findings that the context in which stimuli are encountered can determine either the synaptic or representational change that might occur.

**From single neurons to populations**

While the foregoing is concerned with individual neurons, we know that cognitive representations are driven by *populations* of neurons. It is true that individual cells carry a great deal of information — a simple cell may encode primarily horizontal, or vertical lines — but many of these simple cells converge on hyper-complex cells that may encode much more sophisticated visual stimuli, like the face of one’s grandmother or a particular celebrity (i.e., a grandmother cell, or a ‘Halle Berry’ cell; Quiroga et al., 2005). While this ‘Grandmother Cell Theory’ has not been outright disproven, the advent of multivariate methods has led it to gradually fall out of favour (Barwich 2019).
Indeed, although much is known about the changes in strength at individual synapses and clusters of neurons, it is unlikely that the medium of object or concept representation is the single neuron. Rather, it is likely that the brain’s ability to perceive complex objects in feature space is a result of many units in a neuron assembly activating in tandem to interpret that object based on its simple constituent features (distributed coding theory: DCT; Rumelhart et al., 1986, Polyn et al., 2005). For example, assume we are looking at two objects: a red hardcover book, and a red wooden chair. These objects share many features (e.g., redness and corners), but we can still differentiate between them, because they differ in size, composition and affordances. According to DCT, each unit in a cell assembly is partially responsible for coding a feature of the object in question (Stefanini et al., 2020), like its colour, elasticity, size, texture, etc. However, it is the pattern of coordinated activation among all the units of the cell assembly that allows us to retrieve the object’s identity, in this case being ‘book’ or ‘chair’ (Fig. 1).

The principles of distributed representation scale up to the firing of entire populations of neurons, often in distant physical locations within the brain, as measured by functional magnetic

**Figure 1 Feature-Dependent Co-activation**

A) Activation pattern when a chair is presented. B) Activation pattern when a book is presented. C) Activation pattern when both a chair and a book are presented. Overlapping activational nodes denote shared features. 

*Note.* Model of the neural representational overlap between objects that share features.
resonance imaging (fMRI; e.g., Polyn et al., 2005, Johnson et al., 2009). The complex
conjunctions of features that comprise object concepts are stored in distributed representations,
meaning that different objects can use and reuse the same underlying neural tissue, leading to
partially overlapping memories (e.g., Chadwick et al., 2011, Wimber et al., 2015). Empirical
support for this overlap comes from behavioural evidence that learned associations with reward
can spread to overlapping and adjacent memories through shared connections, even if those
proximal memories have not themselves been conditioned to predict a reward (Ballard et al.,
2019, Wimmer & Shohamy, 2012). It is important to note that these distributed representations
are not limited to a single sensory modality or type of information. Rather, they tend to encode
the ‘idea’ of something. For instance, it has been demonstrated that the same spatial patterns of
voxel activity in occipital cortex are present during both the recognition (seeing) and production
(drawing) of the images they respectively encode (Fan et al., 2019). Thus, these large-scale
activations are not simply visual-evoked responses to objects in feature space, but they seem to encode conceptual representations, or memories, of those objects.

**Representational Change**

It is clear that there is a powerful link between conjunctive neural activation patterns and
memory representations – strong activation can lead to strengthening and moderate activation to
weakening. Critically though, the principles established here also logically extend to the
mechanisms through which multiple memories interface with one another. Analogous to how
high shared activation among individual neurons leads their connections to be strengthened
(LTP), high shared activation, or overlap, of distributed representations will coactivate their
overlapping connections, strengthening their shared features, and causing the representations to integrate or come together in representational space. Conversely, moderate shared activation of
representations will weaken their shared features, causing them to differentiate, or move farther apart in representational space (Detre et al., 2013, Newman & Norman 2010, Ritvo et al., 2019). This is analogous to a scaled-up version of the decoupling that occurs in LTD at the level of individual synapse. The relationship between representation integration/differentiation and co-activation, which can be interpreted as an older cousin to LTP/LTD, often forms a U-shaped pattern known as the non-monotonic plasticity hypothesis (NMPH; Ritvo et al., 2019, Fig. 2; left side).

According to NMPH, one context where integration should occur is when representations strongly or frequently overlap in visual features, or in spatial or temporal context (Lewis-Peacock & Norman 2014, Schapiro et al., 2012, Wammes et al., 2022), while differentiation should occur when this overlap is only moderate (Ritvo et al., 2019, Brunec et al., 2020). Evidence suggestive of this pattern has been observed in several fMRI studies. For example,

**Figure 2 NMPH versus CD**

*Note. The non-monotonic plasticity hypothesis suggests a U-shaped relationship between representational co-activation and integration (solid, purple), while competitive differentiation suggests a negative linear pattern between the two (solid, yellow). The present study aims to reconcile these two hypotheses by comprehensively testing the association between co-activation and representational change. One possibility is that the pattern observed in competitive differentiation constitutes the downward slope of the non-monotonic plasticity pattern (dashed, red), and that the ostensible contradiction between the two hypotheses is a result of limited generalizability across previous literature in the field.*
when pairs of computer-generated fractal images are embedded in a statistical learning paradigm, such that the presentation of one pair item is followed by the other 100% of the time (and therefore, the two memories are likely co-active) their respective representations in the entorhinal cortex (ERC) and in the hippocampus integrated with one another (Schapiro et al., 2012). In contrast, images that were paired only 30% of the time differentiated from one another in those same brain regions. Coincident with these changes in the brain, there was behavioural evidence that strongly paired fractals (which had integrated) were more confusable with one another, whereas weakly paired fractals were less confusable and more individually identifiable. Similarly, visually dissimilar pairs of fractals showed evidence of integration, while moderately similar (i.e., colour-inverted) fractal pairs showed evidence of differentiation. This provided preliminary evidence that learning in a temporal context corresponds to changes both at the neural representational level and the cognitive-behavioural memory level, that are consistent with NMPH, as a function of both temporal co-occurrence, and of visual similarity.

However, the foregoing study included only two levels of temporal or visual overlap, and therefore cannot address the entire possible range of co-activation or overlap, nor could it effectively approximate a U-shaped function. To adequately quantify overlap as a factor driving representational change, more possible levels of co-activation would be necessary. This was accomplished in a later statistical learning experiment, where pairs of images were included that spanned eight possible degrees of visual similarity (0% to 100%) as a proxy for representational overlap. Representational similarity analysis in this study revealed that moderately similar pairs (i.e., moderate co-activation) differentiated in hippocampus, while highly similar pairs (i.e., high co-activation) integrated (Wammes et al., 2022). Converging with Schapiro et al.’s (2012) findings, this study is consistent with NMPH at the neural level. However, there was no
behavioural marker for learning-related change included. While other studies have provided behavioural evidence of either integration or differentiation (Newman & Norman 2010, Keresztes & Racsmány 2012), a behavioural task that can index subtle representational changes in both directions has been elusive. Accordingly, one of the goals of the present study is to interrogate the association between overlap (i.e., visual similarity) and representational change using an approach similar to the foregoing, coupled with a novel behavioural readout of representational change in memory.

**Competitive Differentiation**

While strong imaging and behavioural evidence for hippocampal NMPH exists, an alternative hypothesis suggests a negative linear relationship between the degree of representational co-activation and integration (Fig. 2; right side). Pattern analysis in fMRI has shown that the encoding of overlapping scene-face associations drives differentiation between their respective hippocampal representations, which also reduces later memory interference (Favila et al., 2016). Because the authors suggest that this phenomenon is a result of inter-memory competition for representational space, going forward, we refer to this mechanism as ‘competitive differentiation’ (CD). In another example from the same group, spatial routes that shared greater overlap exhibited greater differentiation of their respective hippocampal and parahippocampal representations (Chanales et al., 2017). CD has also been observed outside of the hippocampus; following an associative memory task, representations of similar colours in parietal and occipital cortex differentiated from one another when colour was the sole metric by which the associations could be distinguished, resulting in lower interference and exaggerated contrast between memories of objects that were highly similar in colour (Chanales et al., 2021,
Zhao et al., 2021). Together, these findings suggest a pattern inconsistent with NPMH, where greater overlap elicits greater differentiation – not integration - between two memories.

**Reconciliation**

While there is some nuance, NMPH and CD make contradictory predictions on the pattern on integration and differentiation of related representations of objects in feature space, where NMPH predicts a U-shaped relationship between co-activation and integration, while CD predicts a negative linear relationship at the highest levels of similarity. To solve this riddle, we need to consider the relative demands of episodic and statistical learning, as they align with the differences in representational activation across the hippocampal anatomy. The hippocampus can be divided into several subfields, each of which has a different profile and preferentially serves different functions. For instance, one subfield called the dentate gyrus (DG) tends to encode sparse representations, meaning that only a small subset of individual units are activated at a given time, and is implicated in memory discrimination and interference reduction (Duncan & Schlichting, 2018). On the other hand, another subfield called CA1 tends to encode more overlapping representations, engages in rapid remapping, and signals the clashes between one’s predictions about the world and the reality of the input from the actual world (Duncan & Schlichting, 2018, Bein et al., 2020). Critical to the current work, integration and differentiation have been shown to take place differentially throughout these hippocampal subfields in a manner that is consistent with the activation properties outlined above. In particular, representations of objects that were strongly reactivated (either due to association or shared episodic context) simultaneously integrated in the CA1 and differentiated in the CA3/DG (Dimsdale-Zucker et al., 2018, Molitor et al., 2021).
These subfields are also part of different pathways within the medial temporal lobe, that are thought to be disproportionately engaged for different tasks and types of learning. The monosynaptic pathway (MSP) is comprised of the direct reciprocal connection between the CA1 and the ERC, while the trisynaptic pathway (TSP) is comprised of the connection of the ERC to the DG, then to the CA3, to the CA1 and finally back to the ERC. According to biologically plausible computational modelling work (Schapiro et al., 2017), the MSP is biased toward statistical learning, or implicit, subconscious learning through cues in the environment which reliably predict event sequences, while the TSP is biased toward episodic learning, or explicit learning of events that can be recalled in distinct, conscious episodes. This is consistent with previous literature which suggests that infants up to 8 months old may excel at statistical learning, but not at episodic learning, due to the slower development of the TSP (i.e., relative to MSP; Saffran et al., 1996, Gomez & Edgin, 2016). The differential engagement of the two pathways in their respectively preferred learning styles may reflect the hippocampus’ adaptation to different learning demands; statistical learning relies on our ability to extract regularities from the environment, which would implicate the liberally overlapping activity patterns of the CA1 in the MSP, while episodic learning requires clear delineation between distinct memories, implicating the orthogonality of the DG in TSP (Sherman et al., 2020).

It is possible that the relative recruitment of the MSP and TSP may evoke different patterns of integration and differentiation, to optimize memory formation in accordance with current goals and task demands (Duncan & Schlichting, 2018). In particular, the type of learning-related representational change that takes place may be dependent on whether information is learned episodically (where competitive differentiation may be necessary) or statistically (where integration and NMPH may be more common).
The Present Work

While previous studies have substantially contributed to our understanding of neuroplasticity and representational change, studies operationalize the similarity, co-activation, or overlap of stimuli in different ways. For example, some studies manipulated visual similarity along a single dimension (e.g., colour in Chanales et al., 2021; shape in Molitor et al., 2021), while others used more complex criteria (e.g., colour, shape, and spatiotemporal position in Dimsdale-Zucker et al., 2018). These inconsistencies complicate the synthesis of general principles across previous empirical work. When it comes to reconciling CD and NMPH, it is possible that the linear CD hypothesis simply constitutes the downward slope of the U-shaped NMPH (Fig. 2, red dashed lines), and that CD’s predicted differentiation pattern already aligns perfectly well with NMPH. Alternatively, it is possible that CD can only be observed in high-demand episodic learning contexts. To our knowledge, no previous literature has explored the differences in representational change between statistical and episodic learning in the context of NMPH or CD, either behaviourally or neurally, despite their proposed differential recruitment of hippocampal pathways. Therefore, the purpose of the present thesis is to (1) explore a novel task as a potential behavioural index of representational change, (2) characterize the nature of representational change along the entire spectrum of visual overlap, and (3) explore how task context interacts with the effect of visual overlap on representational change. To accomplish this, we used visual features from a convolutional neural network (CNN) to generate pairs of novel abstract images that covered the entire span of visual overlap. We embedded these images in both a statistical and an episodic learning task and used a newly developed four-alternative forced choice (4AFC) task to probe for later representational change.
Experiment 1

As outlined above, the goal of this thesis is to characterize how representational change varies as a function of both the degree of overlap between memories, and the type of learning being engaged. In this first experiment, we developed a task designed to provide a generalizable index of representational change. We then tested this task as an outcome measure, both as a function of memory co-activation (operationalized here as visual similarity between paired stimuli) and learning context (operationalized here as episodic versus statistical learning task demands). We generated pairs of images that ranged in visual similarity using a model-based image pair synthesis approach (Wammes et al., 2022), and embedded them in a sequence of images that was shown to participants. Within the sequences, one member of a pair was always presented immediately following its pairmate. This was either known to the participant, creating a more explicit Episodic learning condition, or unknown to the participant, creating a more implicit Statistical learning condition. We were able to read out the representational change that had taken place using a test where participants were briefly cued with an image of Pairmate A, followed by a 4AFC task where they were asked to select the most likely complementary Pairmate B. Unbeknownst to the participants, none of the selectable images was the correct answer, and instead these images were either more or less visually similar to Pairmate A, relative to pairmate B. The participants’ selection during this task indicated integration or differentiation, respectively. Here, we aimed to quantify the relative magnitude of the ‘shift’ in memory either toward or away from the pairmate, to delineate how these representational changes are impacted by initial co-activation and learning demands. Learning-related change consistent with NMPH has often been observed in statistical learning (Schapiro et al., 2012; Wammes et al., 2022), while change consistent with CD has more frequently been observed in episodic learning.
Following from this pattern, we hypothesized that the behavioral magnitude shifts would follow a pattern consistent with NMPH in the Statistical condition, but that in the Episodic condition, it may be more consistent with CD.

**Methods**

**Participants**

Participants were 120 undergraduate students (20 male, 99 female, 1 non-binary; 108 right-handed, 9 left-handed, 3 ambidextrous), ranging in age from 18 to 59 ($M = 26.10$, $SD = 9.86$), recruited from the Queen’s University Psychology Participant Pool. These participants were randomly assigned via Qualtrics to one of our two experimental Learning conditions, Episodic or Statistical. We aimed to collect data until there were 60 usable participants in each Learning condition after exclusions (see Results).

If there was a disproportionate number of participants completing the experiment in one condition while data collection was ongoing, we biased the randomization (through Qualtrics) toward the condition with fewer participants to ensure a balanced $N$ across both groups.

**Materials**

To effectively characterize the nature of representational change, we required pairs of abstract images that sampled the entire spectrum of co-activation/overlap. To achieve this, we aimed to use visual feature similarity as a proxy for overlap. Image pairs were created using an algorithm (Wammes et al., 2022) based on a pretrained CNN architecture (Google’s Inception v3; Szegedy et al., 2015). The network was pretrained to categorize pictures from ImageNet, an online repository containing millions of images of common objects and animals (Deng et al., 2009), and therefore encodes a rich set of visual features. This network takes in image files as three-dimensional arrays of pixels, and runs various small ‘kernels’ or ‘filters’ incrementally
over each position in the image. Each filter responds preferentially to a particular feature in the pixels (e.g., a vertical line, or a dark center and light surround). As the filter moves over the image, it quantifies the relative expression of that feature at each position, thereby creating a spatial map of ‘activation’. Each spatial map comprises a channel in a layer of the neural network. Running many filters over each image creates a stack of activation patterns for many feature channels in every layer, which then is fed forward as the input to the next layer. This next layer will have its own set of increasingly complex filters to run over the input. In early layers, these channels are sensitive to lower-level features, but in later layers, the features increase in complexity such that one channel may preferentially encode objects that look like tropical birds, while another encodes objects that look like houses. In other words, the activations in these neural networks generally represent a hierarchical quantitative description of the visual features present in a given image.

This computation can be then inverted. That is, instead of an image being fed through and eliciting its own characteristic patterns of activation, we can input desired patterns of activation, and iteratively generate an image that elicits them. This is a method often used as a means of visualizing the features that a given channel is sensitive to. For example, one can choose a channel, and inject its preferential features into an image, enhancing and exaggerating features that the channels were trained to detect (Mordvintsev et al., 2015). For example, if we input an image of a tree into the network, it may return filled with features ambiguously resembling birds or houses, assuming the appropriate channels are targeted. Importantly for our purposes, instead of injecting arbitrary features, we can control the balance of channel activity in the network, such that it returns pairs of images that correspond precisely to a researcher-prescribed correlation between visual features in a targeted layer. Consistent with the previous example, we
can control the network to produce image pairs including either two images with roughly equivalent expression of ‘bird’ and ‘house’ features and a feature correlation of 1, or one image with only ‘bird’ features, and one image with only ‘house’ features and a feature correlation of 0. These are the most extreme examples, but this same approach can produce any relative weighting of features in between as well.

Using custom Python (Van Rossum & Drake, 2009) and TensorFlow (Abadi et al., 2016) code, our image pairs were created as follows (Fig. 3). The process begins with a pair of ‘featureless’ greyscale visual noise arrays, allowing the images that are eventually generated to be completely driven by channel activity rather than dependent on pre-existing features. This ‘featureless’ image was deliberately used to minimize existing semantic information and associations for the participant, making the image features relatively arbitrary. Five levels of visual similarity (i.e., feature correlation) were selected ($r = 0.06, 0.28, 0.50, 0.72, 0.94$), such that image pairs ranged from very slightly visually similar (0.06), to almost identical (0.94) (Appendix B), while still leaving space for differentiation (i.e., down to 0.0) and integration (i.e., up to 1.0). We then selected ten sets of two feature channels (e.g., a ‘bird’ channel and a ‘house’ channel); two for each similarity level, such that no individual feature channel was used twice. The algorithm was developed to synthesize image pairs that are interpolated, intermediate images (Fig. 3) between two selected feature channels. For each set and similarity level, two image pairs were synthesized. This created ten total ‘AB’ image pairs, spanning the selected levels of visual similarity. Importantly, this method was also leveraged to generate modified ‘B’ images for each pair, that were either more, or less visually similar to the A image. For example, if an AB pair was synthesized at 0.94 similarity, modified B images could be made at $r = 0.88, 0.91, 0.97$ and
1.00, respectively indicating .06 and .03 away from pairmate A, and .03 and .06 points towards pairmate A in similarity space, relative to the original pairmate B.

**Procedure**

**Learning Task.** The synthesized image pairs were embedded into either a Statistical learning task, or an Episodic learning task, delivered online using lab.js Builder software.
(Henninger et al., 2021). These tasks primarily differed in their instructions, but we will first describe the parameters that were common across the two. After providing some demographic information, participants were shown a continuous stream of images, one at a time. Within these streams, the paired images were embedded such that if the A image from an AB pair was shown, the B image would *always* immediately follow. Within each block, each of the 10 pairs was shown 4 times, with the ordering of pairs randomized. This led to each block containing 40 pair presentations. Participants completed six blocks of this task, leading to 240 total trials for the image presentation sequence.

In all the blocks in both learning tasks, each image appeared on the screen for 1.5 s with a 0.5 s screen containing a fixation cross separating the image screens. In total the image presentation sequence for each of the tasks took 16 min to complete. However, because participants were given a brief, optional intermission between blocks, this phase of the experiment in reality took approximately 20 min.

On 10% of trials, a small grey square would randomly appear, overlaid on the presented image. Participants were instructed to press the spacebar on their keyboard each time they saw a grey square (Appendix C, Appendix D). This was a cover task included in both learning conditions, designed to maintain and monitor participants’ attention to the image sequence while being completely unrelated to the experimental manipulation. In the Statistical condition, this additionally served to obfuscate the true, paired structure of the image sequence. In both conditions, participants were excluded if they performed poorly on this cover task (see Results).

This core task structure was modified slightly depending on the learning condition, keeping as many of the parameters matched as possible (Fig. 4A). In the Statistical condition, participants were instructed to simply attend to the images that appeared on-screen and take in
their detail; importantly, this meant that they were not aware of the pair structure built into the image sequence, a standard precondition for statistical learning (Turk-Browne & Scholl, 2009, Rogers et al., 2021; Appendix C). In contrast, in the Episodic condition participants were informed that the images would appear in pairs, and explicitly asked to memorize them (Appendix D). To further delineate the pairs, the Episodic version displayed a screen saying ‘next pair’ between each pair presentation. All other aspects of the task, stimuli, and timing were identical.

**Four-alternative Forced Choice Task.** In both conditions, after completing the initial learning task, participants were presented with a series of instructions about how to complete the
upcoming (4AFC) task, including a practice trial that used unrelated images that were not used in the rest of the experiment. Following this practice trial and the presentation of instructions, participants went on to complete the actual sequence of 4AFC trials for which experimental data were recorded.

On each trial of this task, participants viewed a brief (1.5s) presentation of the first image from a pair (pairmate A) as a cue, followed by a 4AFC containing four altered versions of the second image in the pair (pairmate B, Fig. 4B). These altered ‘B’ images were designed such that their visual features made them more, or less visually similar to the A image, either by 3 or 6 similarity points. Critically, the correct image was not included among the response options, creating a situation where the participant had to choose an image that was shifted either toward, or away from the cue, allowing the measurement of subtle shifts in underlying memory representations. For example, if pairmates A and B were assigned to be correlated with one another at $r = .50$ during their initial presentation, the 4AFC screen would contain altered versions of pairmate B that were correlated with pairmate A at .56, .53, .47 and .44. These four response options were placed randomly in the top left, top right, bottom, left, and bottom right of the screen, and participants made their choice by clicking on their selected image. While receiving instructions, participants were informed that the correct image may not always be present among the response options, but that they were tasked with choosing which of the four images they believed was most closely resembled the corresponding pairmate. The 4AFC response options remained on screen until the participant provided a response.

Within each block, each of the 10 AB pairs was tested 6 times for a subtotal of 60 trials. There were 4 total blocks yielding a total of 240 trials, and participants were offered a brief, optional break after each block. Because there were multiple selections made for every AB pair,
an aggregate score could be computed for the magnitude shift either toward (i.e., integration), or away (i.e., differentiation) from the pairmate A. If participants were simply randomly responding and there was no systematic bias toward integration or differentiation, then the aggregate scores would indicate no shift. After completing the task, participants were debriefed, and their participation credit was granted.

**Analysis Approach**

To scaffold the presentation of results, the following is a summary of the general approach we take to analyzing the data. First, we report the procedure used to exclude participants’ data from the final sample. This is done based on whether participants completed the task, their performance on our cover task during learning, and in some cases, extraneous participants who exceeded our pre-registered sample sizes were removed.

After implementing our exclusion criteria for each experiment, we analyze the main effects of intra-pair similarity and learning condition on the representational shift of paired memories (measured with 4AFC response), comparing the polynomial patterns these data conform to. This is completed in each condition (first Episodic, then Statistical) using (1) mixed effects modelling to compare linear, quadratic, and cubic model fit, and (2) a cross-validated bootstrap resampling procedure to appraise the generalizability of the observed patterns to new data. Critically, the NMPH predicts a quadratic, U-shaped pattern, where differentiation occurs at moderate similarity, or, potentially, a cubic pattern where little to no representational change takes place at low similarity, differentiation takes place at moderate similarity, and integration takes place at high similarity. Alternatively, CD predicts a negative linear relationship between similarity and representational change. In other words, greater similarity would drive greater differentiation between learned pairmates.
Following this examination of the main effects, we perform (1) contrasts of the representational shifts at each similarity level against zero, to determine whether there is reliable integration or differentiation, (2) contrasts between representational shifts at different similarity levels within each learning condition, and (3) contrasts between representational shifts at the same similarity levels across learning conditions, to determine whether the magnitude shifts differ depending on learning conditions.

**Exclusions**

We first excluded participants with incomplete data (i.e., participants who did not finish either the cover task or the 4AFC task). We then went on to exclude participants who performed poorly on our cover task (i.e., identifying the grey squares), and thus were unlikely to be paying attention or engaging with the task. To do this, we computed their cover task accuracy score (hit rate minus false alarm rate). Participants who scored 0 or less were immediately excluded from the final analysis, as in most cases this meant they had not pressed a button during the entire encoding phase.

Next, we compiled accuracy data for the participants in each condition (Episodic and Statistical). For each condition, we fit the accuracy data to a half-gaussian distribution, and excluded accuracy scores that fell below $2 \times \sigma$ below $\mu$, where $\sigma$ is the measure of spread of the distribution, and $\mu$ is its peak. The half-gaussian was a more sensible distribution (i.e., relative to a full gaussian) as the majority of the distribution was at or close to 100%, with a long tail on the left side.

Participants whose performance survived these exclusion criteria were eligible for the analysis. However, in the case where the number of remaining participants exceeded the maximum pre-registered sample size, we removed the extraneous participants who completed the
experiment most recently to bring our number of participants down to the designated sample size.

Exclusions brought us from $N = 167$ (88 Episodic, 79 Statistical) to a usable $N = 124$ (64 Episodic, 60 Statistical), which were further reduced to $N = 120$ (60 Episodic, 60 Statistical) in accordance with our pre-registered maximum of $N = 60$ per learning condition.

**Results and Discussion**

**Episodic Learning**

**Mixed Effects Model.** To analyze these data, we effect-coded every response in the 4AFC task as either .06 toward, .03 toward, .03 away, or .06 away from the pairmate.

The 4AFC response data were fit to a group of candidate 2-level generalized mixed effects models with different degrees of polynomials (i.e., linear, quadratic, cubic) using the lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) packages in R (R Core Team, 2022, RStudio Team, 2022). In the linear model, Similarity Level was included as a fixed factor and participant intercepts were included as random factors to predict the independent variable, which is the Magnitude of ‘toward’ versus ‘away’ responses. In the quadratic model, in addition to the foregoing, a quadratic term (Similarity Level squared) was included, and in the cubic model, a cubic term (Similarity Level cubed) was added further. As a control, we created an additional null model where no fixed factors were included, and only random intercepts were included. Each model was fit using a Gaussian family, an unstructured covariance matrix, and the Satterthwaite method of estimating degrees of freedom.
After fitting the three candidate models and the control model, we conducted model comparisons by computing Akaike weights \((wAIC)\) and Bayesian weights \((wBIC)\), which are a statistic derived from the log likelihood, that ranges from zero to one, and estimates the probability that a given model is the best fitting of the candidate models (Akaike, 1973, Schwarz, 1978, Vrieze, 2012, Wagenmakers & Farrell, 2004). We also compared the models in order of fit quality using a chi-squared comparison. For the Episodic condition, the cubic model \((wAIC = 1.00, wBIC = 1.00)\) fit better than all candidate models, and fit significantly better than the quadratic model, which was the next best fitting model \((wAIC = .00, wBIC = .00)\), \(\chi^2 (1, N = 60) = 37.61, p < .001\) (Table 1). This indicates that the Similarity Level predicted the Magnitude of

<table>
<thead>
<tr>
<th>Table 1 Experiment 1 Episodic Condition Fit Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Episodic</strong></td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Null model</td>
</tr>
<tr>
<td>Linear model</td>
</tr>
<tr>
<td>Quadratic model</td>
</tr>
<tr>
<td>Cubic model</td>
</tr>
</tbody>
</table>

*\(p \leq .05\); **\(p \leq .01\); ***\(p \leq .001\)

\(\chi^2\) = chi-square statistic; df = degrees of freedom

Note. Fit analytics of each candidate model to the 4AFC response data in the Episodic condition. Each model’s \(\chi^2\) was computed in comparison to the model above it in the table. The bolded row represents the model with the most significant fit.
representational shifts in a cubic pattern partially consistent with the pattern prediction of NMPH (Fig. 5; Episodic). It is only partially consistent because the pattern most commonly associated with the NMPH is driven by the initial flat line when similarity is low, followed by a downward inflection toward differentiation and then an upward inflection toward integration at the higher Similarity Levels (Wammes et al., 2022, see Figure 2; left panel). Our data here show no representational change at the lowest Similarity Level \( r = .06 \), the inflection toward differentiation at a low-medium Similarity Level \( r = .28 \), and the inflection toward integration at a medium Similarity Level \( r = .50 \). However, the cubic pattern appears to be driven by an additional downward inflection that is not predicted by the NMPH, where integration is attenuated after reaching its apex at the highest Similarity Level \( r = .94 \). This is a novel finding that we did not expect to observe, but may be in line with some combination of the influences of

**Figure 5** Experiment 1 Representational Change in Episodic and Statistical Learning

![Graph showing representational change](image)

*Note.* Magnitude of responses shifted ‘toward’ the cued pairmate as a metric for change in representational proximity between the pair items, where positive values mark behavioural integration and negative values mark behavioural differentiation in the Episodic learning group, and Statistical learning group. \( ~p \leq .05 \) (significant before correction); *\( p \leq .01 \); **\( p \leq .002 \); ***\( p \leq .0002 \)
NMPH and CD. We interpret how both mechanisms may jointly contribute to this pattern in the General Discussion section.

**Bootstrapping Model.** To appraise the generalizability of the pattern observed to new data, we employed a cross-validated bootstrapping model-fitting procedure. Using the same 4AFC data as for the mixed effects model, we trained a regression model with the Similarity Level as the independent variable and the Magnitude of ‘toward’ responses as the dependent variable, using the data for all but one held-out participant. This model was then used to make a prediction for the held-out participant, and we computed the Pearson correlation ($r$) between predicted and actual values for that participant. By completing this procedure for every participant, we were left with an $r$ score for each, that indicated the extent to which their behaviour conformed to the remainder of the group’s predicted pattern. Participants’ scores were then resampled with replacement 50000 times, and in each new sample, the average was calculated. In this manner, we computed a non-parametric $t$-test, deriving a $p$-value by determining the number of samples for which the average was below zero (indicating the absence of a positive association between predicted and actual data). This process was done to examine the quality of fit to linear, quadratic, and cubic regression models.

Substantiating the findings from the mixed effects modelling, while the Episodic condition data fit linear ($p = .013$, 95% CI [0.005, 0.078]), quadratic ($p = .011$, 95% CI [0.005, 0.079]), and cubic ($p < .001$, 95% CI [0.024, 0.093]) regression models significantly, the fit to the cubic data was the best. This indicates that not only did the cubic pattern reliably fit the sample, but it is also likely to generalize out-of-sample and is not likely to be a product of overfitting.

**Contrasts.** Given that the foregoing analyses have established a stable pattern of magnitude representational change as a function of visual similarity, we now turn to post-hoc
contrasts to understand the nature of the pattern. That is, whether individual Similarity Levels differ from zero, elucidating whether integration or differentiation occurred to a significant degree at each of our five Similarity Levels, and whether individual Similarity Levels differ from each other, to determine what changes are driving the cubic pattern. These were also conducted using a bootstrap resampling approach. Participants were resampled with replacement 50000 times, and within each resample, participant-level 4AFC response means were computed for each of the five Similarity Levels. These resampled response scores of each Similarity Level were analyzed using a non-parametric t-test for two types of contrasts: (1) Contrasting the scores of each Similarity Level relative to zero, using a p-value computed by finding the proportion of resamples where the average fell above (for differentiation) or below (for integration) zero, and (2) Contrasting the scores across Similarity Levels, using a p-value computed by finding the proportion of resamples where the average pairwise difference between Similarity Levels was below zero.

We conducted five contrasts comparing each Similarity Level to zero, using a Bonferroni-corrected alpha threshold of .01 for each contrast (.05/5). At the low-moderate Similarity Level \( (r = .28: M = -0.12, SD = 1.61) \), there was evidence for differentiation, but this did not survive Bonferroni correction, \( p = .05, 95\% \text{ CI} [-0.261, 0.023] \). The moderate \( (r = .50: M = 0.22, SD = 1.61) \), and moderate-high Similarity Levels \( (r = .72: M = 0.19, SD = 1.56) \) showed significant evidence for integration, \( p = .002, 95\% \text{ CI} [0.067, 0.374]; p = .001, 95\% \text{ CI} [0.065, 0.332] \). The lowest \( (r = .06: M = 0.00, SD = 1.57) \), and highest \( (r = .94: M = 0.06, SD = 1.52) \) Similarity Levels did not significantly differ from zero, \( p = .49, 95\% \text{ CI} [-0.154, 0.146]; p = .19, 95\% \text{ CI} [-0.076, 0.201] \).
We also conducted ten contrasts to explore all possible pairwise comparisons between Similarity Levels using a Bonferroni-corrected threshold of .005 for each contrast (.05/10). We found that both moderate \( (r = .50) \) and high-moderate \( (r = .72) \) Similarity Levels showed significantly greater integration than low-moderate Similarity \( (r = .28), p < .001, 95\% \text{ CI} [0.124, 0.562]; p < .001, 95\% \text{ CI} [0.138, 0.490], \) respectively. No other contrasts yielded a significant difference \( (p > .01, \text{ Fig. 6: Episodic Learning}). \)

Together, these results are broadly consistent with the observed shape of the cubic function and partially support NMPH. That is, there was no change at the lowest Similarity Levels, an inflection toward differentiation (though it did not survive correction for multiple comparisons), then an upward inflection toward integration. Inconsistent with NMPH, there was no evidence for any representational change at the highest Similarity Level.

**Statistical Learning**

**Mixed Effects Model.** Using the same mixed effects model-fitting procedure as in our Episodic condition, we found that, in our Statistical condition where participants were not given explicit pair-learning expectations, the quadratic model was the best fitting model

*Note. Differences in mean representational shift between Similarity Levels. Positive values indicate that the x-axis Level showed higher integration than the y-axis Level, and negative values indicate that the x-axis Level showed higher differentiation. *\( p \leq .005; **p \leq .001; ***p \leq .0001 \)
(wAIC = .40, wBIC = .89), and fit the data significantly better than the linear model, which was the next best fitting model (wAIC = .00, wBIC = .03), $\chi^2 (1, N = 60) = 16.37, p < .001$. Although the cubic model provided a reasonable fit (wAIC = .60, wBIC = .03), it did not significantly improve fit over the quadratic model, $\chi^2 (1, N = 60) = 2.84, p = .092$ (Table 2). This fit, alongside the shape of the pattern of representational change we observed, indicates that Similarity Level is a significant predictor of 4AFC response in a quadratic fashion consistent with NMPH, which predicts a dip into differentiation at moderate similarity ($r = .50$) and integration at the high similarity ($r = .94$). (Fig. 5; Statistical).

### Table 2 Experiment 1 Statistical Condition Fit Quality

<table>
<thead>
<tr>
<th>Statistical</th>
<th>wAIC</th>
<th>wBIC</th>
<th>$\chi^2$</th>
<th>p-value</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null model</td>
<td>.00</td>
<td>.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear model</td>
<td>.00</td>
<td>.03</td>
<td>8.75**</td>
<td>.003</td>
<td>1</td>
</tr>
<tr>
<td><strong>Quadratic model</strong></td>
<td>.40</td>
<td>.89</td>
<td><strong>16.37</strong>*</td>
<td>&lt; .001</td>
<td>1</td>
</tr>
<tr>
<td>Cubic model</td>
<td>.60</td>
<td>.03</td>
<td>2.84</td>
<td>.092</td>
<td>1</td>
</tr>
</tbody>
</table>

*p ≤ .05; **p ≤ .01; ***p ≤ .001

$\chi^2 = $ chi-square statistic; df = degrees of freedom

*Note.* Fit analytics of each candidate model to the 4AFC response data in the Statistical condition. Each model’s $\chi^2$ was computed in comparison to the model above it in the table. The bolded row represents the model with the most significant fit.
**Bootstrapping Model.** Using an identical bootstrapping procedure as the one used for the Episodic condition data, our non-parametric t-test revealed that our data significantly fit a quadratic regression model, $p = .048$, 95% CI [-0.005, 0.066], but not linear, $p = .11$, 95% CI [-0.013, 0.059], or cubic, $p = .11$, 95% CI [-0.012, 0.060]) models, reinforcing the implication of a generalizable quadratic pattern consistent with the predictions of the NMPH, and precluding an overfitting problem.

**Contrasts.** As in the Episodic condition, we turn to post-hoc contrasts to understand the overall pattern. Using five total contrasts comparing each Similarity Level to zero, and a Bonferroni-corrected alpha threshold of .01 (.05/5), we found that only the highest Similarity Level ($r = .94: M = 0.19, SD = 1.61$) resulted in reliable integration, $p = .002$, 95% CI [0.059, 0.325]. While the moderate Similarity Level ($r = .50: M = -0.07, SD = 1.58$) was in the direction of differentiation, none of the remaining Similarity Levels reached significance ($r = .06: M = .05, SD = 1.55: p = .25, 95% CI [-0.089, 0.182]; r = .28: M = 0.08, SD = 1.62: p = .12, 95% CI [-0.054, 0.213]; r = .50: p = .16, 95% CI [-0.197, 0.062]; r = .72: M = 0.07, SD = 1.59: p = .16, 95% CI [-0.077, 0.222]).

We again conducted ten contrasts to explore all possible pairwise comparisons between Similarity Levels in the Statistical condition with a Bonferroni-corrected alpha threshold of .005 (.05/10). We found that high Similarity ($r = .94$) showed significantly greater integration than moderate Similarity ($r = .50$), $p = .001$, 95% CI [0.088, 0.434]. No other contrasts yielded a significant difference ($p > .01$; Fig. 6: Statistical Learning).

**Episodic versus Statistical Learning**

Because we were interested in differential learning patterns across our conditions, we also compared 4AFC response data directly between our Episodic and Statistical conditions.
Contrasts. To delineate differences between learning conditions, we conducted bootstrap resampled pairwise tests analogous to those completed within each learning condition, but here, we instead contrast the scores of each Similarity Level against the scores of their counterpart from the other learning condition. We compute a p-value by finding the proportion of resamples for which the mean difference between learning conditions was below zero.

Using five total contrasts and Bonferroni-corrected alpha threshold of .01 (.05/5), we found that at the low-moderate Similarity Level ($r = .28$) Episodic was more biased toward differentiation than Statistical learning, but this comparison did not survive Bonferroni correction, $p = .02$, 95% CI [0.003, 0.400]. At moderate similarity ($r = .50$), Statistical was more biased toward differentiation than Episodic learning, $p = .003$, 95% CI [-0.491, -0.079]. All other Similarity Levels failed to reach significance ($r = .06$: $p = .34$, 95% CI -0.172, 0.269); $r = .72$: $p = .08$, 95% CI [-0.288, 0.046]; $r = .94$: $p = 0.09$, 95% CI [-0.059, 0.316].

Together, these findings provide an interesting perspective on the differential location of the dip that represents differentiation. It appears that, in the Episodic condition, this dip occurs at a lower Similarity Level ($r = .28$) relative to the Statistical condition, where it occurs at $r = .50$. Following from this, the Episodic condition reaches peak integration at a moderate Similarity Level ($r = .50$), while peak integration does not occur until the highest Similarity Level ($r = .94$) in the Statistical condition, indicating a leftward shift in the Episodic pattern relative to the Statistical pattern. This could indicate that in general, Episodic learning leads to stronger encoding and therefore stronger co-activation, and that as a result, the characteristic non-monotonic course of representational change occurs at lower Similarity Levels in the Episodic condition.
Our findings here establish that, following statistical learning, representational change takes a pattern that is broadly consistent with the prediction of the NMPH, while episodic learning follows a more nuanced pattern, which appears to follow a more compressed non-monotonic shape before exhibiting a strong attenuation of integration at high similarity. We reason that this novel finding may implicate a combinatorial effect of the NMPH and CD when co-activation is high (General Discussion). Additionally, pairwise comparisons revealed that, although the integration effect was significant in both learning conditions, the differentiation effect was not. Despite this, the dip in representational change associated with differentiation contributed to either a reliable cubic or quadratic pattern, which were comparable to the NMPH. This highlights the importance of broadly sampling levels of similarity in order to reliably observe the holistic patterns of representational change. In other words, looking too closely at a constrained range of similarity may prevent one from seeing the overall shape of representational change as a function of shared activation.

It is important to remember that the preceding analyses assume that participants have an established memory of pairmate B, and that their responses on the 4AFC task indicate a distortion of that memory. However, our preceding analyses cannot confirm this assumption, because the trials do not include the correct answer among the response options, and therefore do not allow for an explicit measure for memory accuracy. Because of this property of our task, it is possible that if participants had formed a precise memory for the correct pairmate, then they only chose these 4AFC responses as a “second-best” solution, because they were not given the means to select the actual correct pairmate B. If so, our results might not indicate true underlying representational change, but rather some spurious outcome of random guessing. Indeed, because none of our 4AFC trials so far included the correct pairmate B as a selectable option, our data do
not explicitly allow for the possibility that no representational change had occurred. Moreover, because the Episodic pattern was somewhat divergent from the predictions of NMPH and CD, it was important to further validate that this pattern was a stable and replicable property of learning. Our second experiment was designed to solve both of these issues.

**Experiment 2**

The aims of this experiment were to ensure that participants could actually form a reliable memory for the pairs in the task by including the correct answer among the 4AFC options, and to replicate the patterns of representational change observed in the Episodic condition in Experiment 1. This design will inform us about the stability and accuracy of the established pairmate B memory, while also substantiating our novel findings on representational change. In particular, we can observe whether our previously characterized patterns of representational change still take place when the correct answer (denoting no representational change) is offered. As a secondary benefit, this design would also us to determine whether accuracy differed across similarity levels. One possibility is that because differentiation takes place and could lead to a more distinctive memory trace, accuracy might be improved at moderate levels of initial similarity (where differentiation has been most likely).

Here, we ran two experiments that were almost identical to the structure of the Episodic version of Experiment 1, but with the critical distinction that some or all of the 4AFC trials presented offered a possible correct answer. In particular, for Experiment 2A, one of every six 4AFC trials for each tested pair contained the correct pairmate B as a possible answer. This is in contrast with Experiment 1, where the correct answer was never among the response options; all other aspects of the task or stimuli remained unchanged. For Experiment 2B, all the 4AFC trials presented contained the correct pairmate among the response options. Additionally, we
introduced more variability in the images that were presented. That is, in Experiment 1, though we had two different pairs at each similarity level (and these did not differ from one another in 4AFC performance\(^1\)), these pairs were the same across participants. For example, CNN channel pairs 21-52 and 111-62 were restricted to only create image pairs at \(r = .28\) in Experiment 1 and Experiment 2A, potentially confounding our 4AFC results with features native to the channel pairs, rather than their respective similarity level. To address this, in Experiment 2B, we made each of our 10 CNN channel pairs generate images across all five similarity levels, equally distributed across five experiment versions (Appendix E).

**Methods**

**Participants**

Participants were 120 undergraduate students (17 male, 102 female, 1 non-binary; 107 right-handed, 13 left-handed), ranging in age from 18 to 49 (\(M = 19.30, SD = 3.86\)), recruited from the Queen’s University Psychology Participant Pool.

For Experiment 2B, participants were assigned pseudo-randomly across the five experimental versions via Qualtrics, to ensure there were exactly 12 participants in each version.

**Materials**

**Experiment 2A.** The same stimulus set used in Experiment 1 was used for this experiment.

---

\(^1\)Welch’s two-sample t-test was used to contrast the subject-level mean 4AFC responses the two image pairs unique to each of the five similarity levels. Using a Bonferroni-corrected alpha threshold of .01 (.05/5), we found that, within each similarity level, 4AFC responses did not significantly differ between image pairs \(r = .06: t(124) = 0.38, p = .74; r = .28: t(124) = -1.17, p = .25; r = .50: t(124) = 1.28, p = .20; r = .72: t(124) = -1.27, p = .21; r = .94: t(122) = -0.93, p = .35\).
**Experiment 2B.** To equally distribute the CNN channel-specific features across all five versions of the experiment, we modified our original image pair generation algorithm such that every channel pair generated an image pair for each of our five similarity levels (Appendix E). For example, channel pair 21-52 was made to generate image pairs at \( r = .06, r = .28, r = .50, r = .72 \) and \( r = .94 \), instead of just \( r = .28 \).

**Procedure**

**Experiment 2A.** The experimental procedure was almost identical to that used in Experiment 1’s Episodic condition, with the exception that one of every six of the 240 4AFC trials for every image pair contained the correct answer (i.e., the true, unaltered pairmate B). This meant that eight of the 48 trials for each pair contained the correct answer.

**Experiment 2B.** Participants in this procedure were divided across the five versions of the experiment. Each version tested the same similarity levels as Experiment 1, but with a different set of CNN channel pairs for each. For the 4AFC portion of the experiment, participants underwent only 80 4AFC trials (eight for each tested image pair), all of which contained a possible correct answer, matching the number of trials where the correct answer was possible in Experiment 2A.

**Analysis Approach**

First, we will report the results of our exclusion procedure, which was identical to the one used for Experiment 1.

For each experiment, we will analyze the main effects of visual Similarity Level on Memory Accuracy where the 4AFC trials contained a possible correct answer, and on representational shifts where the 4AFC trials did not contain a correct answer. Similar to our analysis in Experiment 1, there were two general approaches to understanding overall patterns.
across similarity levels. We used (1) mixed effects modelling to assess the effects of Similarity Level on accuracy, and to conduct polynomial model comparison to assess the pattern of representational shifts and (2) a cross-validated bootstrap resampling procedure to assess the generality of the pattern of representational shifts on trials where a correct answer was impossible. In Experiment 2B, we also included experimental Version as a fixed effect in our mixed effects models.

Following this examination of the main effects, in trials where the correct answer was possible, we conducted (1) contrasts of the accuracy scores against chance (0.25) to determine whether there was a reliable memory signal at each Similarity Level, and (2) contrasts between accuracy scores at different Similarity Levels. In trials where the correct answer was not possible, the contrasts were the same as in Experiment 1 – we contrasted representational shifts in each Similarity Level against zero, and then contrasted the shifts between Similarity Levels.

**Exclusions**

**Experiment 2A.** Our exclusion procedure was nearly identical to the one previously outlined for Experiment 1, and brought us from $N = 107$ to $N = 72$. To satisfy our pre-registered max $N$, we removed the last 12 participants from the usable sample, for a final $N = 60$.

Due to a coding error, some of the trials in the 4AFC task where the correct answer was supposed to be present did not actually display the correct answer to the participant. This was specific to a subset of the trials at $r = .72$ and $r = .94$. This error affected just 120 of the 14400 4AFC trials (0.83%). To amend this, we recoded the erroneous trials post hoc to be included in our analysis for representational shifts (i.e., the analysis that assumed a correct answer was impossible), instead of our analysis for Accuracy (where a correct answer was possible).
Experiment 2B. Exclusions brought us from $N = 92$ to $N = 62$. As before, we excluded the last 2 participants for a final $N = 60$. This left us with 12 participants in each of the five versions of the study.

Results and Discussion

Accuracy

Mixed Effects Model.

Experiment 2A. Relevant data points for this analysis were selected by taking only the 4AFC trials where a correct answer for the selection of pairmate B was possible (i.e., the unaltered pairmate B, as it was presented during learning). We first fit a 2-level generalized linear mixed effects model to predict Memory Accuracy (correct: 1; incorrect: 0), with Similarity Level as a fixed factor and including participants’ intercepts as random factors, using a binomial family. As a control, we also fit a null model that included only the random intercepts. Here, the linear model ($wAIC = .70$, $wBIC = .12$) did not fit significantly better than the null model ($wAIC = .30$, $wBIC = .88$), $\chi^2 (1, N = 60) = 3.72$, $p = .054$.

This indicates that Similarity Level had no significant effect on Memory Accuracy, even when controlling for inter-participant intercept variability. That is, participants did not perform significantly better or worse based on the Similarity Level of the image pair being tested.

Experiment 2B. In this experiment, all of the trials included the correct answer, so all of the data were used in the analysis (i.e., no selection of relevant data points was necessary). The analysis was otherwise identical to that of Experiment 2A. Our linear model ($wAIC = .54$, $wBIC = .06$) did not fit the 4AFC data significantly better than the null model ($wAIC = .46$, $wBIC = .94$), $\chi^2 (1, N = 60) = 2.35$, $p = .12$. 
This indicates that, as in Experiment 2A, Similarity Level had no significant linear effect on Memory Accuracy, meaning participants did not reliably perform better or worse based on the Similarity Level being factorized.

**Bootstrapping Model.**

**Experiment 2A.** To test whether accuracy was reliably above chance, we conducted a bootstrapping procedure, similar to the one used in Experiment 1, where participant-level Accuracy scores (aggregated across Similarity Levels) were resampled with replacement 50000 times. Results indicated that Accuracy scores were not significantly greater than chance (.25), $p = .061$, 95% CI [0.244, 0.309], indicating that this learning paradigm, where only one of every six 4AFC trials contained a correct answer, did not result in significant pair memory.

**Experiment 2B.** Here, using the same approach, we determined that aggregated Accuracy scores were significantly greater than chance (.25), $p < .001$, 95% CI [0.475, 0.598], indicating that in the learning paradigm where all presented 4AFC trials contained a selectable correct answer, pair memory was significantly accurate.

**Contrasts.** Taken together, the analyses thus far tell us that there were no differences in Accuracy across Similarity Levels in both Experiment 2A and 2B. Moreover, it appears that Accuracy was not reliably above chance in Experiment 2A, but was significantly above chance in E2B. Consistent with our analysis plan, we carried on conducting post hoc contrasts to determine whether Accuracy was above chance at each individual Similarity Level, and whether Accuracy in any of the Similarity Levels differed from one another.

**Experiment 2A.** To determine whether 4AFC Accuracy at each individual Similarity Level was above chance (.25), we conducted five contrasts using a Bonferroni-corrected alpha threshold of .01 (.05/5). While Accuracy at low-moderate ($r = .28$) and moderate ($r = .50$)
Similarity Levels were above chance, they did not survive Bonferroni correction ($r = .28$: $p = .029$, 95% CI [0.248, 0.373]; $r = .50$: $p = .033$, 95% CI [0.246, 0.356]). None of the other Similarity Levels were above chance ($r = .06$: $p = .19$, 95% CI [0.221, 0.327]; $r = .72$: $p = .37$, 95% CI [0.189, 0.297]; $r = .94$: $p = .30$, 95% CI [0.186, 0.289]; Fig. 7: Some Correct).

These results suggest that participants did not have a precise memory of pairmate B. Interestingly, where there was modest evidence for above-chance memory, it happened to take place at moderate Similarity Levels ($r = .28$ and $r = .50$), where both Experiment 1 and the NMPH predict differentiation will occur. Accordingly, these results may be consistent with previous literature which posits that memories with moderate co-activation become more crystallized and less confusable with each other (Schapiro et al., 2012), potentially as a result of hippocampal differentiation.

Figure 7 Experiment 2 Similarity Level Accuracy Compared to Chance

Note. Mean 4AFC response accuracy on trials where a correct answer was possible when trials were interleaved amongst 4AFC trials where a correct answer was impossible (Experiment 2A, Some Correct) and only trials where a correct answer was possible were shown to the participant (Experiment 2B, All Correct). Dashed lines represent chance (.25). *$p \leq .05$ (significant before correction); **$p \leq .01$; ***$p \leq .002$; ***$p \leq .0002$
**Experiment 2B.** Using five contrasts and a Bonferroni-corrected alpha threshold of .01 (.05/5). Accuracy was significantly above chance at all Similarity Levels ($r = .06; p < .001, 95\% CI [0.404, 0.556]$; $r = .28; p < .001, 95\% CI [0.475, 0.633]$; $r = .50; p < .001, 95\% CI [0.500, 0.658]$; $r = .72; p < .001, 95\% CI [0.438, 0.602]$; $r = .94; p < .001, 95\% CI [0.477, 0.617]$).

Here, in contrast with Experiment 2A, our learning paradigm exhibited accurate pair memories (Fig. 7; All Correct).

Experiment 2 overall indicated that memory performance was significantly above chance when *all* the 4AFC trials contained a possible correct answer, but not when these 4AFC trials were interposed among trials where a correct answer was impossible. Because the 4AFC trials closely followed the image pair presentation sequence during the learning phase, new encoding likely took place during those 4AFC trials where a correct answer was *not* possible. This could have influenced one’s memory in two ways. First, the additional 4AFC trials might have led to further representational change (beyond what occurred during the learning phase), pushing representations even farther from the correct answer. Second, the additional trials *without* the correct answer may have interfered with the active consolidation of the recently learned correct image pairs, modifying the participants’ original memory of the correct pair structure, and precipitating errors during actual pair retrieval (Robertson, 2012).
Representational Shifts

Relevant data points for this analysis were selected by taking the 4AFC trials from Experiment 2A where a correct answer was impossible (five out of every six trials). The remaining data were structurally identical to the Episodic condition of Experiment 1, making this essentially a replication analysis.

Mixed Effects Model. Just as in Experiment 1, we used a mixed effects model fitting procedure to assess the fit of three candidate polynomial models and one null control model; our cubic model fit the data significantly better than all other candidate models ($wAIC = 1.00$, $wBIC = 1.00$), and fit better than the quadratic model, which was the next best fitting model ($wAIC = .00$, $wBIC = .00$), $\chi^2 (1, N = 60) = 36.22, p < .001$ (Table 3).

<table>
<thead>
<tr>
<th>Episodic</th>
<th>$wAIC$</th>
<th>$wBIC$</th>
<th>$\chi^2$</th>
<th>p-value</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null model</td>
<td>.00</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear model</td>
<td>.00</td>
<td>.00</td>
<td>41.30***</td>
<td>&lt; .001</td>
<td>1</td>
</tr>
<tr>
<td>Quadratic model</td>
<td>.00</td>
<td>.00</td>
<td>1.40</td>
<td>.24</td>
<td>1</td>
</tr>
<tr>
<td>Cubic model</td>
<td>1.00</td>
<td>1.00</td>
<td>36.22***</td>
<td>&lt; .001</td>
<td>1</td>
</tr>
</tbody>
</table>

* $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

$\chi^2$ = chi-square statistic; df = degrees of freedom

Note. Fit analytics of each candidate model to the 4AFC response data in the Episodic condition, on trials where a correct answer was impossible. Each model’s $\chi^2$ was computed in comparison to the model above it in the table. The bolded row represents the model with the most significant fit.
**Bootstrapping Model.** To appraise the generalizability of the pattern observed to new data, we used a cross-validated bootstrapping procedure identical to the one used in Experiment 1, and found that the linear, quadratic, and cubic models fit the 4AFC response data significantly ($p = .004$, 95% CI [0.015, 0.100]; $p = .012$, 95% CI [0.007, 0.095]; $p = .003$, 95% CI [0.019, 0.117] respectively). As in Experiment 1, the fit to the cubic data was the best. This indicates that the cubic pattern reliably fit the held-out data, and therefore is likely to generalize out-of-sample.

Remarkably, these representational change results from both mixed effects and bootstrap modelling are largely consistent with those of the Episodic condition of Experiment 1 (Fig. 8), indicating that the unique cubic pattern of representational shifts observed in the Episodic learning condition of Experiment 1 remains when the correct answer is sometimes included among the response options. This highlights that the effect of similarity level on representational shifts is consistent in a second sample, and that it does not rely on the absence of a possible correct answer. In other words, representational shift does not only take place as a “second-best” solution to selecting the correct response. Rather, this shift takes place independently of accurate pair-association learning, and, at least when learned in an episodic context, the representations of paired stimuli follow a polynomial pattern of

**Figure 8** Experiment 2 Representational Change in Episodic Learning

*Note.* Magnitude of ‘toward’ responses as a function of Similarity Level in Experiment 2A, on trials where a correct answer was impossible. ~$p \leq .05$ (significant before correction); *$p \leq .01$; **$p \leq .002$; ***$p \leq .0002$
integration/differentiation based on their visual similarities.

**Contrasts.** To find the size of the effect of representational change at each Similarity Level, we conducted a bootstrapped non-parametric pairwise test identical to the one employed in Experiment 1. Using a Bonferroni-corrected alpha threshold of .01 (.05/5), we found that at low-moderate Similarity ($r = .28$), there was significant differentiation, $p = .002$, 95% CI [-0.431, -0.078], while at moderate similarity ($r = .50$), there was significant integration, $p = .001$, 95% CI [0.106, 0.470]. At moderate-high ($r = .72$) and high similarity ($r = .94$), there was evidence for integration, but neither survived Bonferroni correction ($r = .72$: $p = .040$, 95% CI [-0.014, 0.286]; $r = .94$: $p = 0.021$, 95% CI [0.006, 0.287]). There was no evidence for representational change at low similarity ($r = .06$), $p = 0.46$, 95% CI [-0.142, 0.158]. While there were subtle fluctuations (e.g., reliable differentiation at low-moderate Similarity; sub-threshold integration at moderate-high and high Similarity), the results and the shape of the function were nearly completely consistent with Experiment 1.

To compare representational change between Similarity Levels, we conducted ten contrasts with a Bonferroni-corrected alpha threshold of .005 (.05/10). We found that moderate Similarity ($r = .50$) showed significantly greater integration than low-moderate Similarity ($r = .28$), $p < .001$, 95% CI [0.256, 0.833], as did moderate-high ($r = .72$, $p < .001$, 95% CI [0.130, 0.649]) and high Similarity ($r = .94$, $p < .001$, 95% CI [0.167, 0.624]). No other contrasts reached significance ($p > .01$; Fig. 9).

The purpose of Experiment 2 was to validate our findings from Experiment 1 by eliminating the possibility that the 4AFC data we collected were simply a product of participant acquiescence to the 4AFC task demands. Since none of the 4AFC trials in Experiment 1 actually
contained the correct pairmate B, we suspected that the participants might have selected the “second-best” answer on each trial in the absence of a selectable correct answer.

In Experiment 2A, we repeated the Episodic condition of Experiment 1, but one of every six 4AFC trials contained the actual, unaltered pairmate B as an option. Interestingly, we found that, when correct answers were interposed among our 4AFC trials in Experiment 2A, participants did not reliably select them. Inversely, on trials from the same sequence where a correct answer was not possible, participants selected altered B images in a pattern that was broadly consistent with our findings from Experiment 1. Specifically, representational change took a cubic pattern where differentiation occurs at low-moderate similarity (r = .28), integration occurs at moderate similarity (r = .50), and the integration effect becomes attenuated at high-moderate (r = .72) and high similarity (r = .94). In fact, low-moderate similarity differentiation survived Bonferroni correction in Experiment 2A, whereas it did not in Experiment 1. This indicates that the novel cubic pattern we observed in episodic learning can be replicated across samples.

Figure 9 Experiment 2 Episodic Contrasts of Representational Shift

Note. Differences in mean representational shift between Similarity Levels. Positive values indicate that the x-axis Level showed higher integration than the y-axis Level, and negative values indicate that the x-axis Level showed higher differentiation. *p ≤ .05; **p ≤ .001; ***p ≤ .0001
In Experiment 2B, we aimed to assess the veracity of our learning task, we presented *only* 4AFC trials where a correct answer was possible, and aimed to evaluate accuracy alone, rather than representational change. We found that Memory Accuracy was reliably above chance across all Similarity Levels, which stands in stark contrast to Experiment 2A, where Memory Accuracy was not significant in any Similarity Level. It is possible that accurate pair memory consolidation in Experiment 2A was interrupted by the presentation of trials where a correct answer was absent (Robertson, 2012).

So far, our analyses have coarsely sampled the entire range of intra-pair similarity/co-activation. While this benefits our characterization of overall patterns of representational change, it comes at the cost of precision. Particularly, we found the most evidence for differentiation at low-moderate Similarity ($r = .28$) in Episodic learning, and at moderate Similarity ($r = .50$) in Statistical learning, but cannot pin down the precise inflection point where differentiation stops, and integration begins. In Experiment 3, we aimed to provide a more precise measure of this point, and to determine whether the pattern of representational change within a *constrained* range can be accommodated within the pattern of representational change across the *entire* range.

**Experiment 3**

In Experiment 1, we found that in both versions, the data follow a pattern consistent with the NMPH, with the exception that in the Episodic condition, high visual similarity (i.e., shared activation; overlap) the integration effect was mitigated, and was trending back toward no change. In Experiment 2, we demonstrated that participants can form an accurate memory for the second image in a pair, and we replicated the pattern of representational change in the Episodic condition in Experiment 1. Importantly, across experiments, differentiation was only observed at the low-to-moderate or moderate Similarity Levels, and the point at which differentiation most
strongly occurred was different between learning conditions. To better understand this critical inflection point, we aimed to densely sample the low-to-moderate similarity range, more precisely characterizing the similarity points where the switch between integration and differentiation took place in Experiment 1. Because the shifts differed in Similarity Level across Statistical and Episodic learning conditions, we modelled our design after Experiment 1, instead using five Similarity Levels between \( r = .10 \) and \( r = .70 \), and embedding these in both learning conditions. Like in Experiment 2B, we introduced additional variance in the image sets, such that while the same channel pairs were used to generate images, each CNN channel pair could be assigned to different Similarity Levels depending on the experimental version (Appendix F).

**Methods**

**Participants**

Participants were 200 undergraduate students (19 male, 178 female, 2 non-binary, 1 other; 184 right-handed, 14 left-handed, 2 ambidextrous), ranging in age from 18 to 57 (\( M = 20.42, SD = 5.50 \)), recruited from the Queen’s University Psychology Participant Pool. Participants were pseudo-randomly assigned to our Episodic or Statistical Learning conditions via Qualtrics, until we reached 100 participants per condition after exclusions, and 20 participants in each of the 5 possible versions.

**Materials**

We used the same pair image generation procedure as in Experiment 1 to sample a targeted range of Similarity Levels (\( rs = .10, .25, .40, .55, .70 \)), rather than the broad, general range used in Experiment 1 (\( rs = .06, .28, .50, .72, .94 \)). This was done to more precisely characterize the inflection points within our observed patterns for statistical and episodic pair learning.
Additionally, like in Experiment 2B, CNN channel pairs were used to generate image pairs at all of the five experimental Similarity Levels, and these pairs were divided into five versions (Appendix F), to circumvent the potential confounding effects of channel-specific features, rather than co-activation-specific features, on representational change.

Procedure

The experimental procedure employed in Experiment 3 was identical to the one in Experiment 1, with two exceptions. First, the stimuli ranged from .10 to 0.70 similarity (instead of .06 to .94). Second, participants were distributed five different experimental versions in the same manner as in Experiment 2B.

Analysis Approach

The analysis plan for Experiment 3 is identical to Experiment 1, and is briefly reiterated here. After implementing exclusions, main effects across conditions are analyzed using both a mixed effects modelling approach, and a bootstrap resampling approach. Post hoc contrasts are then completed to determine at what Similarity Levels reliable representational change takes place, and whether Similarity Levels differ from one another. Last, Similarity Levels are contrasted between Episodic and Statistical conditions.

Exclusions

Our exclusion procedure reduced this sample from \( N = 256 \) (131 Episodic, 125 Statistical) to a usable \( N = 207 \) (107 Episodic, 100 Statistical), at which point we removed the last seven participants from the sample, for \( N = 200 \) (100 Episodic, 100 Statistical), where we had 20 participants per version in each experiment version.
Results and Discussion

**Episodic Learning**

**Mixed Effects Model.** As in the previous experiments, 4AFC response data were effect-coded as magnitude shifts from pairmate A, and these data were analyzed using the same generalized linear-mixed effects modelling approach, with three candidate polynomial models and a null control model.

In our Episodic Learning condition, the cubic model fit the data the best among the candidate models ($wAIC = 1.00$, $wBIC = .23$), and fit the data significantly better than the quadratic model, which was the next best fitting model ($wAIC = .00$, $wBIC = .00$), $\chi^2(1, N = 100)$.

**Table 4 Experiment 3 Episodic Condition Fit Quality**

<table>
<thead>
<tr>
<th>Episodic</th>
<th>wAIC</th>
<th>wBIC</th>
<th>$\chi^2$</th>
<th>p-value</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null model</td>
<td>.00</td>
<td>.66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear model</td>
<td>.00</td>
<td>.11</td>
<td>6.50*</td>
<td>.011</td>
<td>1</td>
</tr>
<tr>
<td>Quadratic model</td>
<td>.00</td>
<td>.00</td>
<td>2.79</td>
<td>.095</td>
<td>1</td>
</tr>
<tr>
<td><strong>Cubic model</strong></td>
<td><strong>1.00</strong></td>
<td><strong>.23</strong></td>
<td><strong>18.88</strong>*</td>
<td>&lt; .001</td>
<td>1</td>
</tr>
</tbody>
</table>

*p ≤ .05; **p ≤ .01; ***p ≤ .001

$\chi^2$ = chi-square statistic; df = degrees of freedom

*Note. Fit analytics of each candidate model to the 4AFC response data in the Episodic condition. Each model’s $\chi^2$ was computed in comparison to the model above it in the table. The bolded row represents the model with the most significant fit.
Figure 10 Experiment 3 Representational Change in Episodic and Statistical Learning

Note. Magnitude of responses shifted ‘toward’ the cued pairmate as a metric for change in representational proximity between the pair items, where positive values mark behavioural integration and negative values mark behavioural differentiation in the Episodic learning group, and Statistical learning group. ~p ≤ .05 (significant before correction); *p ≤ .01; **p ≤ .002; ***p ≤ .0002

= 18.88, p < .001 (Table 4, Fig. 10; Episodic). When we included a cubic model where Similarity Level interacted with experiment Version, that model (wAIC = 1.00, wBIC = 1.00) fit significantly better than the cubic model without the interaction (wAIC = .00, wBIC = .00), $\chi^2(16, N = 100) = 181.96, p < .001), indicating that there were subtle differences across versions in the pattern of representational change.

**Bootstrapping Model.** Using the same cross-validated bootstrapping procedure from Experiment 1, we found that our linear model did not fit our 4AFC data significantly, $p = .22, 95\%$ CI [-0.022, 0.054], nor did our quadratic model, $p = .51, 95\%$ CI [-0.039, 0.037] or our cubic model, $p = .27, 95\%$ CI [-0.028, 0.052], indicating that, although our cubic model is the best-fitting among our candidate models in the mixed effects analysis, it may fail to generalize adequately out-of-sample. In other words, the cubic model explains the relationship between Similarity Level and representational change in our sample, but this pattern is not particularly
stable or reliable. This contrasts sharply with the Episodic version of Experiment 1 and the trials of Experiment 2A where a correct answer was absent, where the patterns did appear to be generalizable. It may be the case that the pattern of representational change in an episodic context is sensitive to the demands of the task, such that the cubic relationship between co-activation and representational change is dependent on the range of co-activation sampled. This reinforces the importance of carefully selecting, and controlling for, the range of inter-stimulus similarity, when studying learning-related representational change. It is also noteworthy that, despite the out-of-sample generalization not being as strong, the shape is partially consistent with its E1 counterpart. That is, it qualitatively exhibits a U-shaped pattern followed by an initially unpredicted dip back toward differentiation at this highest Similarity Level. We later explore these possibilities and how they may apply to the current body of literature (General Discussion).

**Contrasts.** Using our contrasting procedure from Experiment 1, with five total contrasts comparing each Similarity Level to zero, and a Bonferroni-corrected alpha threshold of .01 (.05/5), we found that none of the Similarity Levels yielded reliable integration or differentiation ($r = .10; p = .17$, 95% CI [-0.066, 0.182]; $r = .25; p = .26$, 95% CI [-0.167, 0.085]; $r = .40; p = .089$, 95% CI [-0.211, 0.039]; $r = .55; p = .24$, 95% CI [-0.081, 0.170]; $r = .70; p = .12$, 95% CI [-0.198, 0.046]). Indeed, the Similarity Level where representational change was closest to significance was $r = .40$, with a trend toward differentiation that was only marginal, even when uncorrected.

To compare Similarity Levels within the Episodic condition, we conducted ten contrasts using a Bonferroni-corrected alpha threshold of .005 (.05/10). We found that no Similarity Level exhibited significantly greater integration or differentiation than any other Similarity Level (Fig. 11; Episodic Learning).
**Statistical Learning**

**Mixed Effects Model.** In our Statistical learning condition, the linear model \((wAIC = .23, \ wBIC = .96)\) fit significantly better than the null model \((wAIC = .00, \ wBIC = .00)\), \(\chi^2(1, \ N = 100) = 24.19, \ p < .001\) (Fig. 10; Statistical). The quadratic model \((wAIC = .53, \ wBIC = .04)\) and cubic model \((wAIC = .24, \ wBIC = .00)\) improved model fit, but not significantly, \(\chi^2(1, \ N = 100) = 3.67, \ p = .055, \chi^2(1, \ N = 100) = 0.37, \ p = .54\), respectively (Table 5). If we included a linear model where Similarity Level interacted with experiment Version, that model \((wAIC = 1.00, \ wBIC = .06)\) fit significantly better than linear model without the interaction \((wAIC = .00, \ wBIC = .94)\), \(\chi^2(8, \ N = 100) = 75.07, \ p < .001\), indicating that there were differences across Versions in the shape of the linear pattern.

**Bootstrapping Model.** We found that the linear model did not fit our 4AFC data significantly, \(p = .067, \ 95\%\ CI [-0.009, 0.066]\), nor did our quadratic model, \(p = .10, \ 95\%\ CI [-0.012, 0.054]\), or our cubic model, \(p = .21, \ 95\%\ CI [-0.020, 0.045]\). While the linear model appeared to generalize the best, none of the models were reliable out-of-sample, in contrast to the quality of the quadratic fit.

**Figure 11** **Experiment 3 Within-Condition Contrasts of Representational Shift**

Note. Differences in mean representational shift between Similarity Levels. Positive values indicate that the x-axis Level showed higher integration than the y-axis Level, and negative values indicate that the x-axis Level showed higher differentiation.

*\(p \leq .005; **p \leq .001; ***p \leq .0001\)
to the 4AFC response data in the Statistical condition in Experiment 1. The modest negative linear relation, however, is somewhat consistent with the general shape of the NMPH. That is, the range between $r = .10$ and $r = .70$ may not yet reach the levels of similarity at which the flip from differentiation toward integration takes place.

We aim to resolve the analytic inconsistencies between our results from E1 and E3 later in this work, exploring qualitative features of each experiment (General Discussion).

**Contrasts.** With five total contrasts comparing each Similarity Level to zero and a Bonferroni-corrected alpha threshold of .01 (.05/5), we found evidence for differentiation at the highest Similarity Level ($r = .70; p = .048, 95\% \text{ CI} [-0.185, 0.015]$), but this effect did not survive

<table>
<thead>
<tr>
<th>Statistical</th>
<th>wAIC</th>
<th>wBIC</th>
<th>$\chi^2$</th>
<th>p-value</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null model</td>
<td>.00</td>
<td>.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear model</td>
<td>.23</td>
<td>.96</td>
<td>24.19***</td>
<td>&lt; .001</td>
<td>1</td>
</tr>
<tr>
<td>Quadratic model</td>
<td>.53</td>
<td>.04</td>
<td>3.67</td>
<td>.055</td>
<td>1</td>
</tr>
<tr>
<td>Cubic model</td>
<td>.24</td>
<td>.00</td>
<td>0.37</td>
<td>.54</td>
<td>1</td>
</tr>
</tbody>
</table>

*p ≤ .05; **p ≤ .01; ***p ≤ .001

$\chi^2 =$ chi-square statistic; df = degrees of freedom

*Note.* Fit analytics of each candidate model to the 4AFC response data in the Statistical condition. Each model’s $\chi^2$ was computed in comparison to the model above it in the table. The bolded row represents the model with the most significant fit.
Bonferroni correction. All other Similarity Levels did not see significant representational change \((r = .10; p = .24, 95\% \text{ CI } [-0.082, 0.172]; r = .25; p = .14, 95\% \text{ CI } [-0.051, 0.181]; r = .40; p = .39, 95\% \text{ CI } [-0.109, 0.145]; r = .55; p = .36, 95\% \text{ CI } [-0.147, 0.099].\)

To compare Similarity Levels within the Statistical condition, we conducted ten contrasts using a Bonferroni-corrected alpha threshold of .005 (.05/10), and found that no Similarity Level exhibited significantly greater integration or differentiation than any other (Fig. 11; Statistical Learning).

**Episodic versus Statistical Learning**

**Contrasts.** Analogous to Experiment 1, we aimed to further the characterize the differences between learning conditions by contrasting the score for representational change at each Similarity Level to their counterpart score from the other learning condition.

With five total contrasts and a Bonferroni-corrected alpha threshold of .01 (.05/5), we found no learning condition-based bias toward either integration or differentiation at any of the Similarity Levels \((r = .10; p = .44, 95\% \text{ CI } [-0.195, 0.169]; r = .25; p = .12, 95\% \text{ CI } [-0.071, 0.283]; r = .40; p = .14, 95\% \text{ CI } [-0.087, 0.298]; r = .55; p = .23, 95\% \text{ CI } [-0.246, 0.111]; r = .70; p = .44, 95\% \text{ CI } [-0.166, 0.146].\)

Our findings here provide an interesting extension to our findings from Experiment 1. Specifically, Statistical learning showed a negative linear relationship between Similarity and representational change, partially consistent with the downward slope of the quadratic pattern found in its counterpart from Experiment 1. Surprisingly, Episodic learning did not show a pattern consistent with the analogous Similarity range in its Experiment 1 counterpart, and instead seemed to take on a cubic pattern of its own that will need to be replicated in later work. With this, Statistical learning seems to have an overall quadratic pattern whose components can
be individually detected by densely sampling narrower ranges of similarity/co-activation (Fig 12; Statistical). In contrast, Episodic learning seems to exhibit a task-dependent pattern that may rely on the range of similarity/co-activation being sampled (Fig 12; Episodic).

**Figure 12 Experiments 1 & 3 Representational Change in Episodic and Statistical Learning**

*Note.* Comparison between representational shifts in Experiment 1 and Experiment 3 across Statistical and Episodic learning conditions. In Statistical learning, the linear pattern observed in Experiment 3 with a constrained range of similarity levels can be accommodated within the quadratic pattern observed in the broad similarity range of Experiment 1, and predicted by the non-monotonic plasticity hypothesis (NMPH). This can be observed when data points are combined across both Experiments into one overall pattern (light grey line). In contrast, the patterns observed in Episodic learning differs across Experiments 1 and 3, indicating that the pattern observed may be dependent on the range of similarity levels present in the task context.

**General Discussion**

The present work aimed to develop a behavioural index for subtle changes in representational organization in memory as a function of learning and stimulus overlap, then to use that method to test theoretical accounts of neuroplasticity across different types of learning (Episodic and Statistical). This follows from prior work that discovered changes in patterns of representational overlap in the hippocampus as a result of statistical learning (Wammes et al.,
but did not identify how this pattern manifested itself behaviourally. To our knowledge, the current study is the first to elucidate this issue. Because different kinds of learning are thought to bias the hippocampus toward particular processing pathways (Schapiro et al., 2017; Gomez & Edgin, 2016), we predicted that the two different learning types may manifest themselves in different pair learning outcomes and different patterns of representational change (Dimsdale-Zucker et al., 2018; Ritvo et al., 2019; Schapiro et al., 2017; Chanales et al., 2021).

Across three experiments, we probed for representational change using a 4AFC task where the correct answer was not among the response options, so that participants were forced to choose an image that skewed either toward (i.e., integration), or away (i.e., differentiation) from its pairmate. We found that while statistical learning was largely consistent with the predictions of the NMPH, episodic learning was more complex.

Specifically, in Experiment 1, we found that statistical learning led to representational change that followed a quadratic pattern, with a dip toward differentiation at moderate levels of similarity, and then integration at the highest levels of similarity. On the other hand, episodic learning followed a cubic pattern not wholly consistent with either NMPH or CD. Representational change seemed to follow a non-monotonic pattern lower in the similarity spectrum in the Episodic condition than in the Statistical condition. That is, both the dip, and the rise toward integration were at lower similarity levels. The Episodic pattern, however, showed an attenuation of the integration effect at the highest similarity levels, which was not predicted by either hypothesis.

In Experiment 2, we further validated the 4AFC task by ensuring the participants were able to choose the correct answer when it was present, and replicated the Episodic pattern uncovered in E1. Here, we discovered that participants formed accurate pair memories, but this
accuracy was undermined when there were interposed trials that did not contain the correct answer. We reasoned that since these trials likely took place within the temporal window for consolidation (Dharani, 2014), they may have caused new encoding, or additional representational change to take place (Robertson, 2012).

Last, we conducted Experiment 3 to explore the differential dips in representational change (denoting differentiation) that occurred in the learning condition of Experiment 1. Through our dense sampling procedure, we found that in Statistical learning, the dip toward differentiation occurred when visual similarity was .70, which plausibly can be accommodated into the non-monotonic pattern observed in Experiment 1. In contrast, while we found that Episodic learning still exhibits a cubic pattern, it was divergent from the nature of the pattern in Experiments 1 and 2, and showed only weak and non-significant differentiation, albeit at a level of visual similarity (0.4) that might be expected based on the preceding experiments. These results were surprising, as they contradicted our initial predictions that the resulting patterns in each learning condition would be internally consistent. While Statistical learning was consistent, at least in part, with prior experiments, Episodic learning was not. We expected the Episodic condition of E3 to follow a quadratic pattern resembling the initial, NMPH-like curve observed in its E1 counterpart, before $r = .94$ where the integration effect was mitigated. We explore possible causes and implications of this contradiction later in this discussion.

**Statistical Learning and the NMPH**

Taken together, our findings suggest that when statistical learning occurs, our behavioural index of representational change as a function of the similarity spectrum followed a pattern consistent with NMPH. In other words, when participants were presented with paired images without their knowledge or any explicit expectation of learning, behavioural
representations of memories of moderate to high intra-pair similarity integrated closer together, while memories that were moderately similar to one another differentiated further apart.

Evidence for this comes from our findings in E1, which found that representational change varied quadratically as a function of similarity level. In this case, the ‘dip’ where representational change becomes negative represents a trend toward differentiation between the learned pairmates. We then attempted to further interrogate the dip in Experiment 3 by densely sampling the moderate similarity range, and found that representational change followed a negative linear relationship, such that it exhibits little to no representational change at low similarity, and slight differentiation at moderate similarity. In this case, the pattern observed in Experiment 3 conceptually ‘fits’ into the pattern found in Experiment 1, consistent with the predictions of the NMPH. This implied that the behavioural data from this task may serve as a useful readout of underlying representational change as a result of learning.

The results here are consistent with previous neural data on similarity-based representational change (Lewis-Peacock & Norman 2014; Schapiro et al., 2012; Wammes et al., 2022). In particular, previous research had shown an association between both visual similarity and temporal co-activation, and representational change, such that both moderate visual similarity and moderate temporal co-occurrence led to differentiation (Schapiro et al., 2012). Our behavioural findings here are consistent with the neural pattern suggested by this work, such that differentiation only occurs at moderate levels of co-activation. As mentioned in the introduction, Schapiro et al. (2012) also discovered that pairmates that neurally differentiated from one another were less confusable with each other in behavioural data, while the opposite effect occurred when pairmates integrated together. As an aside, we found in Experiment 2A that only the moderate similarity levels showed pairmate retrieval that was modestly above chance (though
not significantly). This may suggest a link between representational differentiation of moderately co-active memories and improved memory accuracy.

Research following up on Schapiro et al. (2012), found a neural pattern of representational change that was similarly consistent with NMPH. Specifically, using stimulus base generated in the same manner used in the present thesis, Wammes et al. (2022) showed that the dentate gyrus (DG) subregion of the hippocampus exhibited differentiation between moderately visually similar pairmates and integration between highly similar pairmates. This pattern of neural data, which emerged from using the same task and a similar stimulus generation method, is remarkably consistent with the pattern of behavioral data observed in Experiment 1 here. As mentioned in the introduction, this U-shape is quite common in several behavioural studies (Newman & Norman 2010, Keresztes & Racsmány 2012) and in neural data (Schapiro et al., 2012, Wammes et al., 2022). Together, these results strongly suggest that representational change between associated memories is mirrored in the behavioural data derived from our 4AFC measure of memory and representational change.

Experiment 3 showed a negative linear pattern that also appears as though it could be accommodated with the U-shape observed in Experiment 1. We believe this highlights an important methodological point; had we erroneously described the highest similarity level included in Experiment 3 \( (r = .70) \) as subjectively ‘high similarity’, we might have interpreted our findings as evidence that differentiation occurs at the highest similarity levels, consistent with CD. This reinforces how findings might appear mixed, when contrasts are made between only a handful of ‘overlap’ conditions, and an objective metric of the possible range of similarity or co-activation is not considered. We believe this is evidence that follow-up research should
continue to use objective measures for co-activation as an effector of representational change, such that future findings can be more generalizable within the literature.

**Episodic Learning and Inconsistent Representational Change**

Interestingly, when the same stimuli used for the statistical learning paradigm were encoded episodically, the patterns were less consistent across experiments. In other words, when participants were instructed that the images would be presented in pairs and that they should be memorized, representational change followed a pattern that was not predicted by either NMPH or CD. That is, while NMPH predicts a U-shaped pattern, where differentiation occurs at moderate similarity and integration at high similarity, we observed a cubic pattern, where this characteristic U-shape occurred, but then integration became attenuated at the highest similarity level. When we constrained the range of similarity levels in Experiment 3, none of the patterns of representational change was significant, but, qualitatively, representational change looked to follow a similar trajectory despite the truncated range. That is, while there was an initial dip at low-moderate similarity representing differentiation (and similar to Experiment 1), the highest similarity level also showed slight differentiation, which contradicts NMPH’s prediction of integration at this level. In both cases, the highest similarity levels in their respective experiments skewed more toward differentiation than predicted by NMPH, and than observed in the Statistical condition. This de-integration effect may reflect some influence of CD, which predicts that when representations are highly similar to one another, the memories will repel one another and differentiate (Chanales et al. 2021). This mechanism has been observed in several cases, including learning overlapping face-scene pairs (Favila et al., 2016), face-word pairs (Drascher & Kuhl, 2022), spatial routes (Chanales et al., 2017), and faces paired with similarly colored objects (Chanales et al., 2021).
Hybrid Learning Mechanism in Episodic Learning?

One possibility is that the observed cubic pattern is the result of an amalgamation of the two theories of neuroplasticity (i.e., NMPH and CD). Initially, following Episodic encoding, the pattern seems to reflect NMPH lower in the similarity spectrum until it reaches high similarity, at which point the integration effect becomes attenuated. In previous studies that exhibit the pattern we are labelling CD, differentiation occurs at higher similarity levels, where almost complete representational overlap between memories results in the highest amount of differentiation (Chanales et al., 2017, Chanales et al., 2021). It is possible that when explicitly encoding, where encoding distinct, trial-unique memories is important, we have a mechanism that actively tries to highlight distinguishing features that separate the most similar memories. As a result, our finding of attenuation of integration at the highest similarity levels may be the result of a combined influence of NMPH of CD, where the former is biased towards integration, and the latter towards differentiation (Fig. 13). Namely, toward the highest point in the similarity spectrum, the differentiation predicted by CD manifests itself, imposing an additive influence on the conventional, quadratic pattern of NMPH. This changes the pattern into a cubic one, where representational change trends back towards zero.

However, when explicit coding does not take place, and items do not need to be memorized in strict, trial-unique episodes, the influence of CD seems to disappear, resulting in a conventional quadratic pattern of representational change, canonized by NMPH, suggesting a highly task-specific nature for the manifestation of CD.

Because differentiation was exhibited at different points between Experiment 1 and Experiment 3, it is possible that CD is also range-dependent, where the relative similarity range being sampled may cause the downward slope of CD to occur earlier in the range (see Figure 13,
That is, the competitive differentiation mechanism might apply not at the highest similarity level *possible*, but to the highest similarity level *present* in the task context. Future research can use predictive modelling to determine whether this is the case.

**Figure 13** *Hybrid Model of Representational Change as a Function of Co-activation*

![Diagram of Hybrid Model](image)

*Note.* Across all experiments, episodic learning manifested a cubic pattern of representational change that is not fully accounted for by either NMPH or CD. We propose a hybrid model where, at high levels of similarity relative to the similarity range being sampled, CD expresses a downward slope that influences the upward slope of NMPH, pulling the curve down to zero, and attenuating the integration effect that is characteristic of NMPH at high similarity levels.

**Differentiation at Different Similarity Levels**

In E1, we found that the dip toward differentiation was located in a different place depending on learning conditions. Specifically, the similarity level that the dip was most prominent at was $r = .50$ in the Statistical condition, while it was at $r = .28$ in the Episodic condition; a clear leftward shift. To understand this finding, it is worth returning to the initial
intention for our design. We used an algorithm that generates paired images with a prescribed similarity level to one another, with the intention of using this as a proxy for co-activation (as in Wammes et al., 2022). However, there are other elements, including encoding strength or prediction, which may drive up co-activation. In the Encoding condition, participants are made explicitly aware that they are to learn pair associations prior to the learning task. With this, and that the design repeats each pair many times, participants were likely primed to make predictions about the B image when the A image was on screen (Kim et al., 2017). This active engagement with the learning task could in turn increase coactivation of the two representations (i.e., relative to the more incidentally encoded Statistical condition, e.g., Bein et al. 2020), meaning that at a given similarity level, the true co-activation value would be higher for episodic than statistical. This in turn predicts that the differentiation and integration predicted by NMPH occur at lower similarity levels in episodic relative to statistical learning. Future research will be needed to fully appraise this possibility, perhaps by titrating the number of repetitions of each pair.

A Possible Model of Learning Change

One of our most surprising findings was the pattern of representational change observed in the Episodic learning condition, where there was a ‘de-integration’ effect at the highest similarity levels, and the dip appeared at lower similarity levels than in Statistical learning. We mentioned previously that this pattern may be the result of the additive influences of both NMPH and CD. To explain our findings here, we propose a model where NMPH applies universally, but some of its integrative effects are offset by CD at the highest similarity levels when learning episodically. At the same time, when learning episodically, the overall strength of co-activation is increased. This in turn leads to the dip occurring earlier in Episodic than in Statistical learning,
because of greater co-activation at lower similarity levels. Of course, future research will be needed to vet and refine this theoretical model on new data.

There remains an open question as to why competitive differentiation would take place only in the episodic condition and not the statistical condition. One interpretation is that the purpose of Episodic encoding is to encode unique, non-overlapping representations of individual memories. Related to this, Episodic learning is biased toward engaging the trisynaptic pathway in the hippocampus, which passes through the highly orthogonal representational structure of DG. This may contribute to less integration and greater differentiation across pairmates in Episodic relative to Statistical learning, as Statistical learning is biased toward engaging the monosynaptic pathway, which involves the more liberally overlapping representations of CA1, favouring integration (Schapiro et al. 2017). In Episodic learning, the TSP may become engaged, initializing a competitive mechanism that causes de-integration of the pairmates at high similarity. Future imaging studies should explore the engagement of TSP and MSP as a result of these different kinds of learning, and how it affects the resulting representational change between paired stimuli in the hippocampus proper.

**Evidence for Memory Despite Representational Shifts**

When 4AFC trials were included containing the correct target among the response options, accuracy was only above chance when only these trials were included and not when they were interleaved among the standard 4AFC trials where a correct response was impossible. Providing compelling evidence for the stability of the patterns of representational change, on those trials where a correct answer was impossible, responses followed a cubic pattern remarkably similar to that observed in our first episodic learning experiment. This indicates that the pattern generalized to a brand-new sample of participants, and that this pattern manifests
even participants have the (occasional) opportunity to select the correct pairmate. The
behavioural outcome of representational similarity may be partially retrieval-dependent, because
participants were able to accurately recall the correct pairmates when they persistently had a
choice to make a correct selection. It is also clear that the pattern in representational shifts occurs
regardless of the veracity of the participant’s memory for the correct pairmate, given that the
pattern still occurs when the correct answer is occasionally included as an option.

Limitations

One major limitation of this study is that there remain some unexplained inconsistencies
across the experiments. When the moderate similarity range was densely sampled, the pattern
observed in the Episodic condition of Experiment 1 did not align with our findings in a
constrained range in Experiment 3, even though all other components of the experiment were
identical. It is possible that if we look too closely at smaller similarity ranges within the greater
pattern (as we did in Experiment 3), it may be difficult to measure patterns of representational
change that can be contextualized by the ideas presented in NMPH. This may help to demystify
the ostensible contradiction among shift patterns in previous literature; studies that find different
patterns of representational shift may be looking at different parts of the same pattern. Indeed,
our goal was to find a generalizable pattern of representational change by sampling the entire
spectrum of possible intra-pair similarities, but this was done at the cost of precision in our
characterization of these patterns. We cannot pinpoint at exactly what percentage of intra-pair
similarity integration switches to differentiation or vice versa. Future work could remedy this by
using more granular measures of image similarity. That is, one could use a comparable image
generation procedure to vary intra-pair similarity across all possible levels from 0-100% rather
than only five levels that coarsely sample this range. More granular measures of representational
change could also be employed. For example, Chanales et al. (2021) used a colour wheel as an outcome task, where participants could freely select colour options from an entire spectrum between pairmates, in contrast to our study, where participants were limited to only four options per trial, leading to less precise observations of representational change.

While we established a behavioural method of detecting representational change that aligns with prior work in statistical learning (Wammes et al., 2022), another limitation is that we do not have any neural data with which to compare or map to the behavioral effect. Thus, we cannot draw a direct link between representational shift at the neural level and at the behavioural level, only infer it using previous literature. This inference has no protection against possible confounding effects of inter-sample and inter-subject variability. It will be important to run additional fMRI studies to observe the representational change occurring at a neural level to verify its consistency with our behavioural findings and validate our task.

Last, although we tested Memory accuracy for pairmates after episodic learning in Experiment 2, we did not measure accuracy following Statistical learning. In other words, we do not have a direct comparison between accuracy in the two learning conditions, and thus cannot identify a clear link between accuracy and representational change, as a function of learning condition. If our suggestion that representational differentiation is associated with increased memory accuracy and decreased pairmate confusability holds true, we predict greater Memory Accuracy at $r = .28$ in statistical learning, given that slight differentiation occurs earlier in the circuit than in episodic learning. Future studies should explore the variation of pairmate retrieval accuracy across learning conditions, and further elucidate the link between accuracy and representational change.
Conclusion

The present work found a potential behavioural index for representational change among paired memories, and characterized a broad pattern of this change among two different major learning paradigms. We found significant evidence for NMPH at the behavioural level in Statistical learning, and some suggestion that Episodic learning might be additionally driven by the partial influence of CD at the highest similarity levels. We believe the present study can help reconcile our understanding of the interaction among these two hypotheses by introducing a combinatory model that accounts for both mechanisms, and opens the door to a holistic conceptualization of representational change.
References


https://doi.org/10.1214/aos/1176344136


Appendix A

Research Ethics Approval

This study was approved by the General Research Ethics Board of Queen’s University (TRAQ# 6029803).
Appendix B

Image Pair Sample

23-108 Pair at .06 Similarity

23-108 Pair at .94 Similarity
Appendix C

Statistical Learning Instructions

Welcome!

Please make sure this window is in fullscreen.

Thank you for participating in our study! In this study, you will view a series of colorful images, one at a time. Your job is to pay close attention to the images, taking in their detail. Occasionally, a small grey square will appear, overlaid on the image. Here is an example of what that might look like:
You are tasked with monitoring for the small grey squares contained in the images. There will never be more than one grey square per image, and some images may not have a grey square at all.

When you find a grey square, press the Space bar.

Note that pressing Space will not cause the screen to disappear, but rest assured we have received your response!

This first task will take about 20 minutes, but you will have the option to take a short break every 2 - 3 minutes. These breaks are included to ensure that when you are completing the actual task, you are giving it your full attention. It is important that you remain focused and provide responses throughout (unless it is during a break), because if you do not provide a response, or respond too slowly for several trials, we will not be able to use your data for our research.

When you’re ready, please press Space to begin.
Appendix D

Episodic Learning Instructions

Welcome!

Please make sure this window is in fullscreen.

Thank you for participating in our study! In this study, you will view a series of colorful images, one at a time. The images will be presented in pairs. That is, each image will always be followed by its pairmate. Pairs will be separated by a screen that says "next pair" in large green text.

Your job is to try and memorize these pairs. Each pair will be repeated many times.

Occasionally, a small grey square will appear, overlaid on the image. Here is an example of what that might look like:
When you find a grey square, press the Space bar.

Note that pressing Space will not cause the screen to disappear, but rest assured we have received your response!

This first task will take about 20 minutes, but you will have the option to take a short break every 2 - 3 minutes. These breaks are included to ensure that when you are completing the actual task, you are giving it your full attention. It is important that you remain focused and provide responses throughout (unless it is during a break), because if you do not provide a response, or respond too slowly for several trials, we will not be able to use your data for our research.

When you're ready, please press Space to begin.
Appendix E

Experiment 2 Paired Channel Distribution Across Similarity Levels

<table>
<thead>
<tr>
<th>Experimental Version</th>
<th>CNN Channel Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>94</td>
</tr>
<tr>
<td>C</td>
<td>72</td>
</tr>
<tr>
<td>D</td>
<td>50</td>
</tr>
<tr>
<td>E</td>
<td>28</td>
</tr>
</tbody>
</table>
Appendix F

Experiment 3 Paired Channel Distribution Across Similarity Levels

<table>
<thead>
<tr>
<th>Experimental Version</th>
<th>CNN Channel Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>89-86</td>
</tr>
<tr>
<td></td>
<td>81-34</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>70</td>
</tr>
<tr>
<td>C</td>
<td>55</td>
</tr>
<tr>
<td>D</td>
<td>40</td>
</tr>
<tr>
<td>E</td>
<td>25</td>
</tr>
</tbody>
</table>