From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Psychology

2022

The Role of Competition in Representational Change

Ritvo, Victoria Jackson-Hanen

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp013n204226m

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

THE ROLE OF COMPETITION
IN REPRESENTATIONAL CHANGE

Victoria J.H. Ritvo

A DISSERTATION PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE DEPARTMENT OF PSYCHOLOGY

Adviser: Ken Norman

May 2022

© Copyright by Victoria J.H. Ritvo, 2022. All rights reserved.

Abstract

What are the principles that govern whether neural representations move apart (differentiate) or together (integrate) as a function of learning? According to supervised learning models that are trained to predict outcomes in the world, integration should occur when two stimuli predict the same outcome. Numerous findings support this, but—paradoxically—some recent fMRI studies have found that pairing different stimuli with the same associate causes differentiation, not integration. The goal of this dissertation is to provide an explanation for these paradoxical results and explore the repercussions of such an explanation. In Chapter 1, I argue that supervised learning needs to be supplemented with unsupervised learning that is driven by spreading activation in a U-shaped way, such that inactive memories are not modified, moderate activation of memories causes weakening (leading to differentiation), and higher activation causes strengthening (leading to integration). In Chapter 2, I present a series of behavioral experiments that provide a roadmap for exploring these effects. In Chapter 3, I present a neural network model that instantiates a U-shaped learning rule, and I show how it can explain a range of findings in the literature. The model can lead to integration or differentiation and makes several novel predictions—for instance, that differentiation is rapid and asymmetric. This work provides a unifying theory underlying a wide array of findings in learning and memory, and it provides a blueprint for future work in the field.

Author's note

The Introduction and Chapter 1 were adapted from: Ritvo V.J.H., Turk-Browne N.B., & Norman K.A. (2019). Nonmonotonic plasticity: How memory retrieval drives learning. *Trends in Cognitive Sciences*. 23(9), 726-742.

Chapter 3 was adapted from a manuscript in preparation: Ritvo V.J.H, Nguyen, A., Turk-Browne, N.B., & Norman, K.A. (2022). Differentiation and integration of competing memories: A neural network model.

Chapter 3 includes links within the text to several videos. The videos can be found at the following locations:

- Video 1: https://www.youtube.com/watch?v=Xx1RWKDnbm0
- Video 2: https://www.youtube.com/watch?v=-9sllcYDdwA
- Video 3: https://www.youtube.com/watch?v=GjRUgZoRX4M
- Video 4: https://www.youtube.com/watch?v=--Bxmsg9VuA
- Video 5: https://www.youtube.com/watch?v=v1TeMSDK1N8
- Video 6: https://www.youtube.com/watch?v=5GUIREJrLMM
- Video 7: https://www.youtube.com/watch?v=u7Uq3OFWgTY
- Video 8: https://www.youtube.com/watch?v=uNDr95wZ-F0
- Video 9: https://www.youtube.com/watch?v=CQc8rVUNzI0

Acknowledgements

There are so many people to thank for their role in this work. First and foremost: my adviser, Ken Norman. It's been a privilege to have you as a mentor and an example, to see your unending curiosity lead your research and help shape my own approach. From the first meeting I had with you to my last, you've prioritized my growth as a scientist. You're also one of the best explainers of complicated things I've ever met. And when my late-husband's health drew me away from the lab, you met my situation with understanding and flexibility, especially when I naively thought I wouldn't need to take a leave of absence: many of the early versions of the models in this dissertation were coded while I sat in chemo rooms. Thank you for not questioning my needs and supporting me in all aspects of my life.

Nick Turk-Browne, thank you for choosing me as your lab manager. One of my first days on the job you told me I was going to run one of the most complicated and ambitious fMRI experiment ever. I was terrified I would fail, but you had no reservations. I have never learned any skill so fast and efficiently in my life, and I wouldn't be the scientist I am now without that experience. Thank you for believing in me, for always knowing when I needed to be hand-held, and for knowing when I needed to be thrown in the deep end.

I'd like to thank the other professors at Princeton I've worked with: Jon Cohen, Yael Niv, and Erik Nook. Thank you for all your input and guidance. My other collaborators: Alex Nguyen most of all, who has been deeply involved in our modeling work, as well as Cătălin Iordan, and my research assistants Daniel Greenidge and Jordan Gunn. Everyone I've worked with on the mental health survey project at Princeton: Abby Novick, Mai Nguyen, Judith Mildner, Sarah Wilterson, Edan Daniel, Dan Mirea. I'm grateful to everybody in the Computational Memory Lab and the Turk-Browne Lab at Princeton. Some people I'd like to thank in particular, either for being early mentors or continued voices of academic or personal support: Megan deBettencourt, Alexa Tompary, Natalia Cordova, Anna Schapiro, Judy Fan, Cameron Ellis, Angela Radulescu, Ben Hutchinson, Sarah DuBrow, Anne Mennen, Nina Rouhani, Rolando Masis-Obando, Sebastian Michelmann, James Antony, and Lizzie McDevitt.

I'm also grateful to all my committee members: Ken and Nick, as well as Adele Goldberg, Brice Kuhl, and Tim Buschman. Thank you for all your guidance. I'd also like to thank the people who read and shared insights about the papers that went into this dissertation: Judy Fan, Chris Honey, Justin Hulbert, Jarrod Lewis-Peacock, Ehren Newman, Jordan Poppenk, Randy O'Reilly, Andrew Saxe, Anna Schapiro, Lizzie McDevitt, and Alex Nguyen.

I'm indebted to the professors I worked with as an undergraduate at Columbia, Peter Balsam and Ed Smith, as well as the graduate students there who taught me how to run experiments, Teal Eich and Heather Van Volkinburg.

I have an incredible community of loved ones who kept my life stitched together when it felt like everything was falling apart. While you may not have been directly involved with this research, it would not be possible without your tremendous love.

Josephine Simple, more than anyone else in the world, you keep me sane. You remind me every day of the best parts of myself. Please keep reminding me, I never tire of it. I don't think it's possible for me to ever get bored if I'm talking to you: It says something that when I'm busy I have to preface our phone calls with "we'll have to keep it short today, I only have two hours today to talk." I know you feel the same way, or you wouldn't have put up with my recounting

every episode of *Survivor* or *Buffy the Vampire Slayer* scene-by-scene. I'm amazed at the endless canal of your enthusiastic ear. You are the definition of chosen family.

Andrew Kahn, pairmate to Josephine in the greatest inheritance I received from Max. He insisted I befriend you after he died, and I'm so glad I did. Your way of needling into my brain always surprises me. You make me think clearer and more compassionately. You've also given me the best COVID lifelines possible—your old apartment, and you as my neighbor. Moving here is the best decision I ever made, and a lot of how I've managed during the pandemic is knowing you're upstairs.

I'm lucky to have a community of friends too large to name here, but some people I have to mention specifically for their help over the last few years: Abby Novick and Ben Hoskin, our Wednesday dinners were the highlight of my week. Thank you for always making me feel like family. Thank you to everyone in His Majesty the Baby (Josephine and Andrew, but also Helen Campbell, Jer Senie, John Griswold, and by extension Jenny Downes), the writer friends I picked up from Max (Carlie Hoffman, Eliza Schrader, Sarah Matthes, Bonnie Antosh, Ava Kofman) and an assortment of other friends: RC Saint-Amour, Daniel Tay, Shelby Lohr, Jenne O'Brien, Matt Powell, B Kelley, Alex Forstenhausler, Hannah Weinberg-Wolf, Ed Su, and Michel Djandji. Henry and Louis Yegerman, you are friends who I can barely call friends because you're actually my family.

I'd like to thank Mrs. Saint Pierre, the first teacher in school who saw my incessant questions not as a sign of academic weakness, but a sign of a budding academic. Thank you for showing me the power in not understanding.

Most of all, my family. My big brother, David Jackson-Hanen: You are my best friend, my third parent, and my OG academic advisor. I think my joy at having you explain things is only matched by your joy in explaining them. When I was in middle school you paid me five pounds to 1) receive a lesson from you on the fundamental theorems of calculus and 2) meet you at the High School the next day so I could recite the theorems by heart to your math teachers. I didn't tell you then but I would have done it for free to make you proud. Thanks for always answering my math questions, from elementary school through graduate school.

My little brothers, Misha and Danya: I love answering your questions, whether about coding in Python or speed-runs in Zelda. You've given me the best example of competition-dependent learning—for years I've loved using your adorable, confusable younger selves as a starting point in explaining this research. Anya, my fourth parent: thank you for being the calm, joyful, rock for all chaotic people around you. I'm grateful that you're in my family.

Mom and Dad, thank you for giving me a researcher's brain. I recommend to anyone to grow up in a family of academics, it's a riot. Dad, thank you for your unwavering support in my toughest times—you are my knight in shining armor. And I love that every week for the past few years you've asked me to update you on my research. Explaining neurofeedback and the Nonmonotonic Plasticity Hypothesis to you has made me a clearer thinker and communicator. Mom: I challenge anyone to find a mom who is more a mom at her core than you. Thank you for always being my champion, and always reminding me there's somewhere to turn when it looks like there's nowhere to go. I am unfathomably lucky for your wisdom, your care, and your cat videos.

Lastly, my love, my shared-brain, my best friend, my late-husband Max Ritvo. He had a hand in much of this work. I know his strong opinions on what to write about in my NSF statement helped me earn the fellowship. But there's so much more than that: Art and science went hand-in-hand in our relationship. Neuroscience trickled into his poems, and metaphor

flowed into the way I talk about my research. One of Max's biggest fears was that my inevitable grief would prevent me from finishing my PhD. But grief is what got me to the finish-line. My grief is an echo of my love for him, and no thoughts ever give me as much support as those of Max. Max has reoriented my life toward love, and without him, I probably wouldn't have thought to make these acknowledgements more than a couple lines. Max, these words are your words.

Contents

Table of Contents

Abstract ..... *iii*

Introduction *1*

Thesis Overview **2**

Chapter 1: Nonmonotonic plasticity: How memory retrieval drives learning *4*

Supervised learning and representational change **4**

Data that challenge supervised learning **7**

Unsupervised learning and representational change **9**

How the NMPH accounts for representational change **12**

Testing the NMPH **19**

What is the function of the NMPH? **20**

Balancing supervised and unsupervised learning **22**

Concluding remarks and future directions **23**

Chapter 2: Studies of how competition affects memory for continuous stimulus values *24*

Introduction **24**

Experiment 1 ..... **27**

Methods 27

Results and discussion 33

Experiment 2 ..... **34**

Methods 35

Results and discussion 35

Experiment 3 ..... **37**

Methods 39

Results and discussion 41

Experiment 4 ..... **42**

Methods 44

Results and discussion 44

Experiment 5 ..... **46**

Methods 46

Results and discussion 47

Experiment 6 ..... **48**

Methods 50

Results and discussion 50

Experiment 7 ..... **52**

Methods 52

Results and discussion 53

Experiment 8 ..... **54**
Methods 55
Results and discussion 56

General discussion **58**
Lessons 58
Implications for future studies 61
Summary 63

Chapter 3: A neural network model of differentiation and integration ***64***

Basic network properties **65**
Model architecture 66
Inhibitory dynamics 68
Learning 69
Competition 69

Model of Chanales et al. (2021): Repulsion and attraction of color memories **69**
Key experimental findings 69
Potential NMPH explanation 70
Model set up 71
Results 75
Important lessons 84

Model of Favila et al. (2016): Similar vs different predictive associations **85**
Key experimental findings 85
Potential NMPH explanation 87
Model set up 87
Results 89
Important lessons 91

Model of Schlichting et al. (2015): Blocked vs interleaved learning **92**
Key experimental findings 92
Potential NMPH explanation 92
Model set up 94
Results 97
Important lessons 99

Discussion **99**
Differentiation requires a rapid learning rate and is sensitive to activation dynamics 100
Asymmetry of representational change 103
Limitations and open questions 105
Summary 107

Methods **108**
Model architecture 108
Learning 111
Stimulus presentation 112
Practical advice for getting the model to show differentiation 113

Conclusion ***117***

References ***120***

Introduction

How do stored memories change as a function of experience? The neural representations of past events can get stronger or weaker individually, but can also change with respect to each other, with neural overlap either decreasing (differentiation) or increasing (integration). These changes to the similarity structure of memories have an enormous effect on subsequent retrieval, affecting how much memories compete (less overlap results in less competition and thus better recall of distinctive features [1]) and how much generalization occurs (more overlap leads to more generalization [2]).

Here we address two fundamental, inter-related questions: (1) What are the learning rules that govern how representations change in the brain? and (2) According to these rules, which situations lead to differentiation vs. integration? These questions have come to the forefront of learning and memory research in recent years, driven by new, multivariate fMRI analysis methods that make it possible to track how neural similarity structure changes with learning [3]. For instance, representational similarity analysis involves comparing the spatial patterns of brain activity evoked by different stimuli or conditions, and can tell us whether stimulus-evoked brain patterns are becoming more or less similar with learning [4]. These new methods have led to a wealth of new fMRI results that have proven to be highly constraining about underlying learning mechanisms. Some of these new results are well-explained by classic supervised learning models, which posit that the brain adjusts representations to predict outcomes in the world [5,6]. Consistent with these theories, such studies find integration when two stimuli predict the same outcome and differentiation when two stimuli predict different outcomes (e.g., [7–9]). However, contrary to these findings, other fMRI studies have found that linking two stimuli to the same associate leads to differentiation rather than integration [10–12].

These studies present a challenge to supervised learning models, which predict that the neural weights of these memories should be adjusted to make them more similar to each other. This dissertation aims to make sense of these seemingly contradictory results.

Thesis Overview

In Chapter 1, I review the studies that pose a challenge to classic supervised learning models. I argue that explaining representational change requires supplementing supervised learning rules with unsupervised learning rules that adjust neural representations based on how strongly memories are activated during the retrieval process. Furthermore, I argue that the function relating memory activation to learning is U-shaped, such that inactive memories are not modified, moderately activated memories are weakened, and highly activated memories are strengthened.

In Chapter 2, I present a series of behavioral experiments designed to investigate how the theory presented in Chapter 1 may lead to distortions in recall. While the studies run did not lead to conclusive findings, they provide insights that can help researchers investigate this topic in the future.

In Chapter 3, I present a neural network model that can explain the findings discussed in Chapter 1. The model uses a purely unsupervised learning algorithm, and can lead to integration or differentiation depending on the amount of activity allowed to spread to competitors—inactive competitors are not modified, moderately active competitors are weakened (leading to differentiation), and highly active competitors are strengthened (leading to integration). The model also makes several novel predictions—most notably, that differentiation requires a large learning rate and is asymmetric. These modeling results provide a computational explanation for

the diverse set of seemingly contradictory empirical findings, as well as new insights into the dynamics at play during learning.

The work in this dissertation contributes to our understanding of how we learn about our world. The existing literature in this area is complex and full of seemingly contradictory results. I provide a unifying theory for these results while also providing a roadmap for future research.

Chapter 1: Nonmonotonic plasticity: How memory retrieval drives learning

Supervised learning and representational change

As noted in the Introduction, a very influential account of representational change is supervised, error-driven learning. According to this view, the goal of learning is to minimize prediction error: the discrepancy between predicted and actual outcomes in the world. This idea has been instantiated in models that apply error-driven learning algorithms (e.g., error backpropagation [5]) to multi-layer ("deep") neural networks. These models have experienced a resurgence of popularity in recent years as a result of successful applications to large-scale problems in AI (e.g., computer vision [13]) and neuroscience (e.g., modeling the ventral visual stream [14,15]). Substantial progress has also been made in considering how these algorithms could be implemented in the brain [16–21].

Supervised learning models adaptively re-represent the input patterns in hidden layers of the network (between the input and output layers) to facilitate mapping from inputs to outputs. Learning rules like backpropagation adjust connection strengths throughout the network to minimize prediction error in the output layer. These adjustments can change the similarity structure of hidden-layer patterns evoked by inputs, pushing them together or pulling them apart. In these models, if two inputs map onto the same output, error-driven learning pushes hidden-layer representations of the inputs together; if the inputs map onto different outputs, error-driven learning pulls their hidden representations apart [6] (**Fig 1.1)**. Importantly, these representational changes are incremental in nature. The effect of supervised learning is to adjust hidden-layer representations just enough to support accurate prediction, which requires sensitivity to both

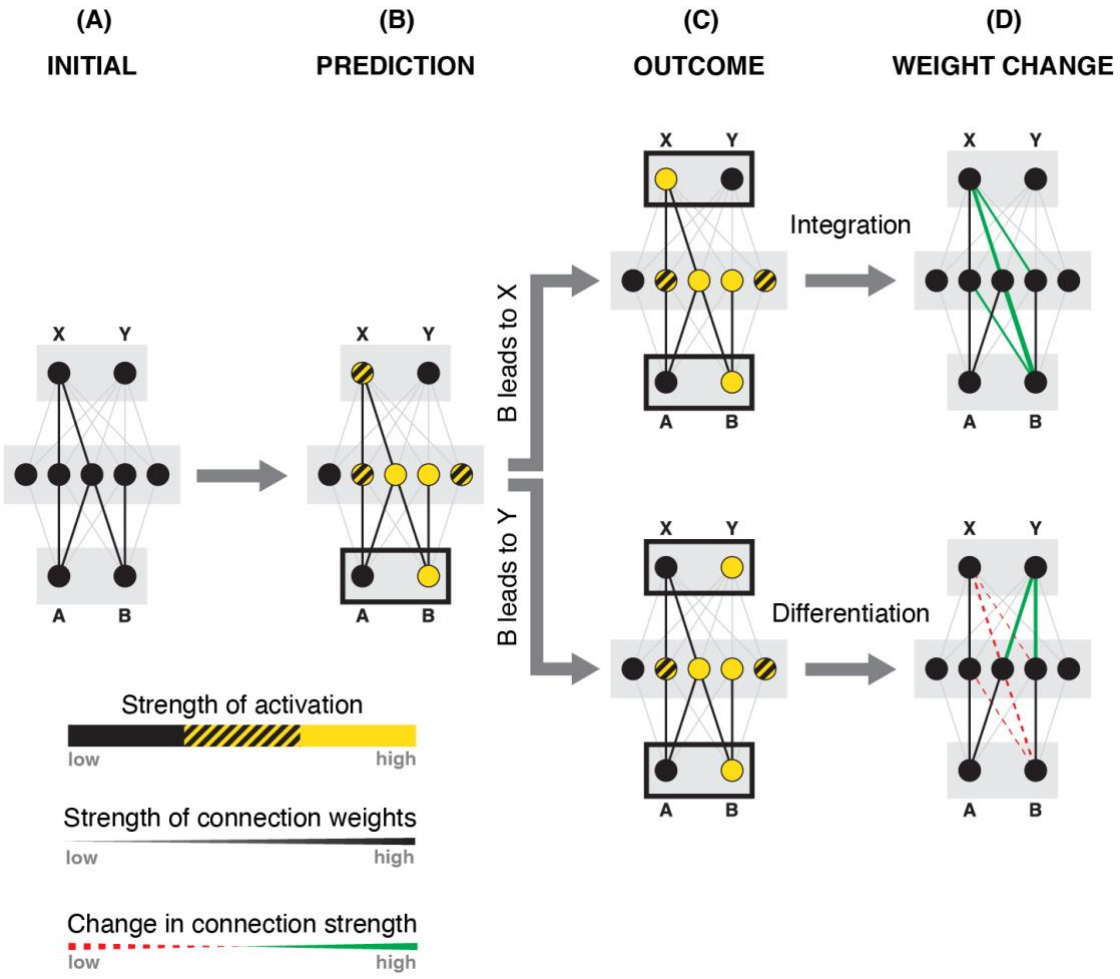


Figure 1.1: How supervised learning leads to differentiation and integration. *Supervised neural network learning algorithms adjust connections between units based on the difference between a prediction phase (where an input is presented to the network and activity spreads to the output layer) and an outcome phase (where the "correct" output is presented to the network).* ***Fig 1.1*** *depicts a three-layer, feedforward network (i.e., activation flows unidirectionally from input to output) that learns.* ***(A)*** *This network has already been trained to map A to X; also, A and B start out with overlapping hidden representations (as indicated by strong connections from both A and B to the middle hidden unit). Now we want to train the network using backpropagation such that either B predicts X (i.e., B and A have the same predictive consequence) or B predicts Y (i.e., B and A have different predictive consequences).* ***(B)*** *When B is presented to the network during the prediction phase, X becomes partially active because B activates some of the same hidden units that A previously activated, which (in turn) had been previously linked to X.* ***(C)*** *During the outcome phase, the correct output (either X or Y) is activated; learning is based on difference between output-layer activity in the prediction and outcome phase. In the B-X case (top), X was activated less in the prediction phase than in the outcome phase. To remedy this (**D**; top), backpropagation strengthens weights from active hidden units to X and also from the B input to the active hidden units that are most strongly connected to X (and thus are in the best position to boost X's activation). The result of this learning is to increase the extent to which A and B project to the same hidden units. In the B-Y case (**C;** bottom), X was activated too strongly in the prediction phase and Y was not strongly activated enough. To remedy this (**D**; bottom), backpropagation strengthens weights from active hidden units to Y, reduces weights from active hidden units to X, and also weakens weights from B to the active hidden units that are most strongly connected to X (and thus are most strongly "responsible" for X being over-activated). The result of this learning is to reduce the extent to which A and B project to the same hidden units. Importantly, these predictions about differentiation and integration are not limited to the feedforward architecture shown here—the predictions generalize to recurrent supervised learning algorithms that allow activation to spread in both directions. Several recurrent algorithms have been shown to mathematically approximate feedforward backpropagation [17,20], and thus the same learning outcomes should occur.*

similarities and differences in what the inputs predict [22].

A key feature of supervised learning is its asymmetry: In these models, learning to predict distinct outputs (e.g., two different-looking faces) from similar inputs (e.g., two similar-looking scenes) yields different representational change effects than the opposite (learning to predict similar scenes from distinct faces). In the former, two similar-looking scenes should be pulled apart; in the latter, two distinct faces should be pulled together. As such, when testing the predictions of supervised learning models on actual experimental data, it is important to identify which part of the experimental design corresponds to the input (i.e., the cue that triggers the prediction) and which part corresponds to the output (i.e., what is being predicted). Sometimes this is straightforward (e.g., in paired-associate learning paradigms where participants are given a cue and asked to recall an associate) and sometimes less so (e.g., when to-be-associated items are presented simultaneously—in this case, participants could implicitly generate predictions by taking either one of the paired associates and trying to retrieve the other).

Numerous fMRI results support the predictions of supervised learning models about representational change: For example, one study [9] found that training participants to predict a shared scene associate from object cues (by making them imagine the object in that scene) results in integration (see also [23,24]). Supervised learning also explains how *community structure* affected neural representations in another study [7]: Participants viewed a sequence of arbitrary symbols, where symbols transitioned to other symbols in the same community more frequently than to symbols in other communities; neural representations of symbols within communities integrated, relative to symbols across communities, in both the cortex and the hippocampus [7,8]. Neural network models using supervised, error-driven learning (where each stimulus was used to predict the next) were able to simulate these results [7,25].

Data that challenge supervised learning

Recently, a series of fMRI studies have challenged the supervised learning account of representational change, by showing that linking memories to the same associate can sometimes lead to differentiation [10–12,26]. For example, when participants were trained to predict the same face in response to two similar scenes (e.g., two pictures of barns), this led to differentiation of the hippocampal representations of the scenes, to the point where the two scenes became less neurally similar to each other than they were to unrelated stimuli (e.g., two barns ended up as less neurally similar than a barn and a bridge) [10]. Intriguingly, this effect may vary across brain regions: One study found that linking two objects to a common associate can lead to differentiation of the objects' representations in right posterior hippocampus and integration in left medial prefrontal cortex [11] **(Fig 1.2)**. Another study looked at the effects of presenting objects in the same episodic context (a video "walkthrough" of a house) vs. a different context, and found integration in one subfield of the hippocampus (CA1) but differentiation in another (CA2/3/DG) [12].

Importantly, these differentiation effects cannot simply be explained in terms of hippocampal pattern separation—the automatic bias for hippocampus to represent stimuli in a relatively orthogonalized way, driven by sparse coding [27–30]. One reason is that differentiation in the above studies does not occur automatically, but rather only in response to the demand of learning a shared associate. Furthermore, as noted in several recent papers [10,31,32], these differentiation effects in some cases go beyond pattern separation: Orthogonalization implies that all items should have the same (low) overlap, but in the study described above [10], visually similar scenes linked to the same face had lower levels of hippocampal pattern similarity than visually dissimilar scenes (see also [31]). Lastly, whereas

differentiation has often been reported in the hippocampus, it has also been observed in other regions (e.g., anterior medial prefrontal cortex and bilateral inferior frontal gyrus [11]).

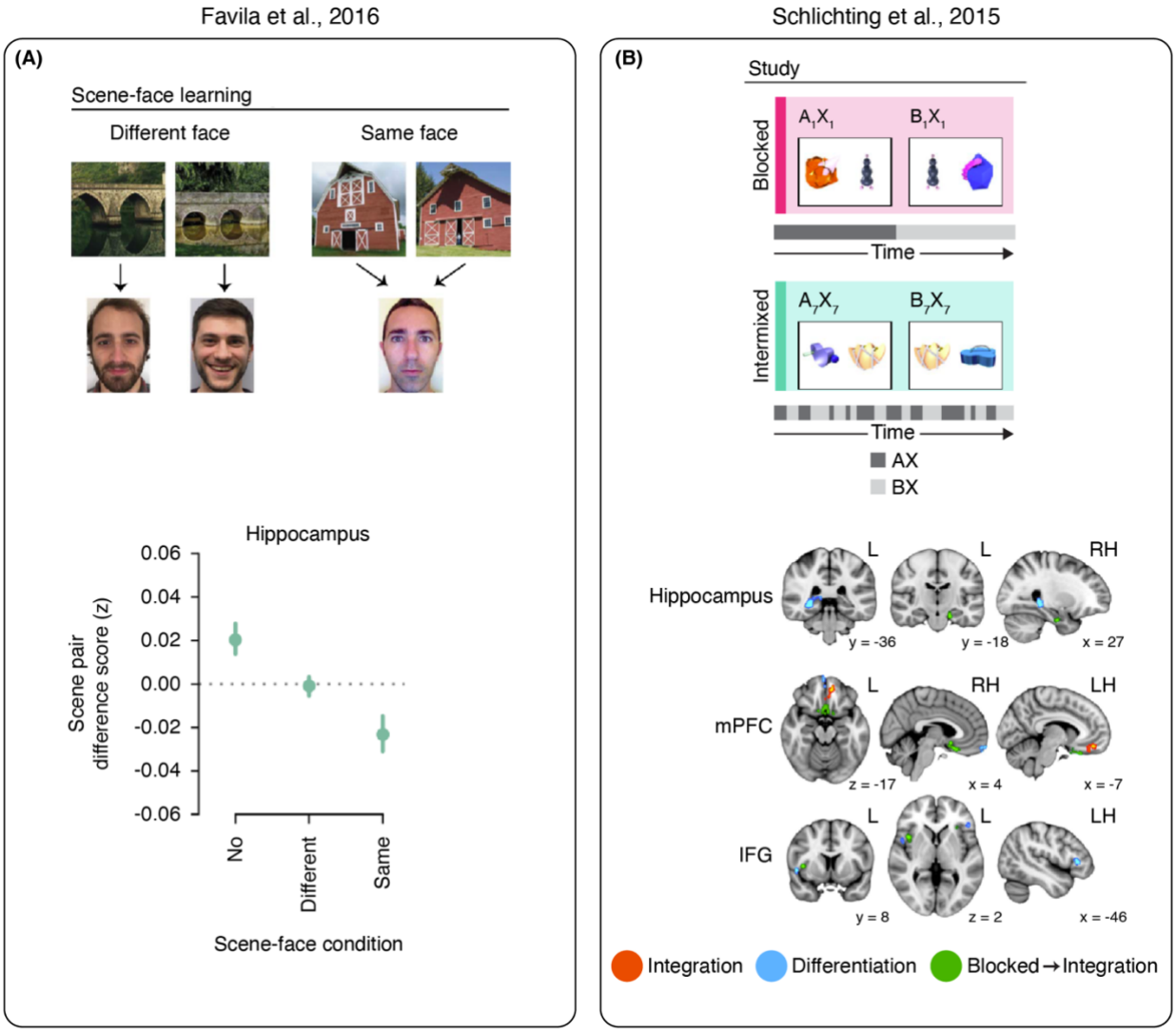


Figure 1.2: Linking stimuli to a shared associate can lead to differentiation. (A) *A recent fMRI study [10] used pairs of similar scenes (e.g., two barns) as stimuli. Participants were shown scenes as cues and trained to predict faces. Specifically, participants associated each scene in a pair with a different face, the same face, or no face. Representational similarity was measured using a scene pair difference score: neural similarity between stimuli in the same pair (e.g., two barns), compared to stimuli in different pairs (e.g., a barn and a bridge). According to supervised learning models, stimuli with similar predictive consequences (i.e., same-face scenes) should integrate [6]. However, in the hippocampus, pairmates in the same-face condition showed differentiation rather than integration. Figure adapted from [10].* ***(B)*** *In a related fMRI study [11], participants encoded object pairs (e.g., AB, BC, DE, EF) such that some objects (A, C) were indirectly linked by a shared associate, B. Some related pairs were learned in a blocked fashion (all AB before any BC) whereas others were learned in an interleaved fashion. The key question was how linking to a shared associate affected the neural similarity of A and C, relative to baseline pairs that were not linked by a shared associate. The analysis focused on hippocampus, medial PFC (mPFC), and inferior frontal gyrus (IFG). Within these regions of interest, some areas (colored blue in the figure) showed differentiation for both blocked and interleaved training (e.g., right posterior hippocampus; note that this differentiation effect was referred to as separation in the original paper). Others (colored red) showed integration for both blocked and interleaved training (e.g., left medial PFC). Still others (colored green) showed integration for blocked training and differentiation for interleaved training (e.g., right anterior hippocampus). Figure adapted from [11],*

The aforementioned studies provide a particularly stark challenge to supervised learning models. However, it is not hard to find other memory results that pose a challenge for these supervised learning models, hiding in plain sight. One example is the testing effect: the finding that successfully retrieving a memory in the absence of feedback leads to superior recall, compared to simply restudying the material [33–37]. For example, say that you learn the word pair "absence-hollow"; later, you are given "absence-" and you retrieve the associated word ("hollow") without feedback. From the perspective of supervised learning, one could interpret retrieval of "hollow" as a prediction that is not matched by the world (because of the lack of feedback), so we might expect weakening of the memory based on this prediction error. Alternatively, one could argue that "hollow" does not really constitute a prediction about what will happen in the world (since participants do not expect feedback), and hence there will be no prediction error and thus no learning. Crucially, neither of these scenarios explains why retrieval of "hollow" is a strongly positive (reinforcing) learning experience, as has been demonstrated by myriad studies [35].

Unsupervised learning and representational change

How can we explain the mix of findings described above? Supervised learning is too useful to discard—the vast majority of the recent successes of neural network research, in both AI and neuroscience, depend on the ability to learn arbitrary input-output mappings via supervised learning [13]. Thus, rather than replacing it, we propose that supervised learning needs to be supplemented with other learning principles.

One particularly relevant set of learning principles is the class of *unsupervised* neural learning rules. These rules adjust synaptic connections throughout a network based on

information that is local to the synapse (typically, the activation of the pre-synaptic and post-synaptic neurons), without any explicit consideration of how well the network is predicting external outcomes. The simplest form of unsupervised learning is the classic Hebbian rule ("fire together wire together"), which specifies that correlated firing of the pre- and post-synaptic neurons leads to synaptic strengthening [38,39]. This rule plays an important role in classic models of hippocampal contributions to memory (e.g., [27]). At the level of neural representations, this Hebbian rule predicts that correlated neural firing will lead to integration [40,41], even in the absence of explicit teaching signals [42–44]. There is substantial empirical support for this prediction. For example, in monkeys, increasing the correlation of inputs from the hand by surgically connecting the skin of two fingers eliminated the discontinuity between neural zones representing adjacent digits, indicating that correlated activity boosts neural similarity [45].

This simple Hebbian learning principle can also explain some findings that are challenging for supervised learning. For example, it can easily explain why retrieval of an association (e.g., retrieving "hollow" from the cue "absence-") reinforces that memory: With Hebbian learning, coactivity of the representations of the cue and the associate leads to strengthening of the connection between them. Importantly, however, Hebbian learning cannot straightforwardly account for the findings reviewed above of increased differentiation for two stimuli linked to a shared associate—this should make the neurons in their representations "fire together" more, leading to integration, not differentiation.

A more sophisticated form of unsupervised learning might do better. In particular, several learning rules have been described with the property that no learning occurs when memories are inactive, weakening occurs when memories are moderately activated, and strengthening occurs

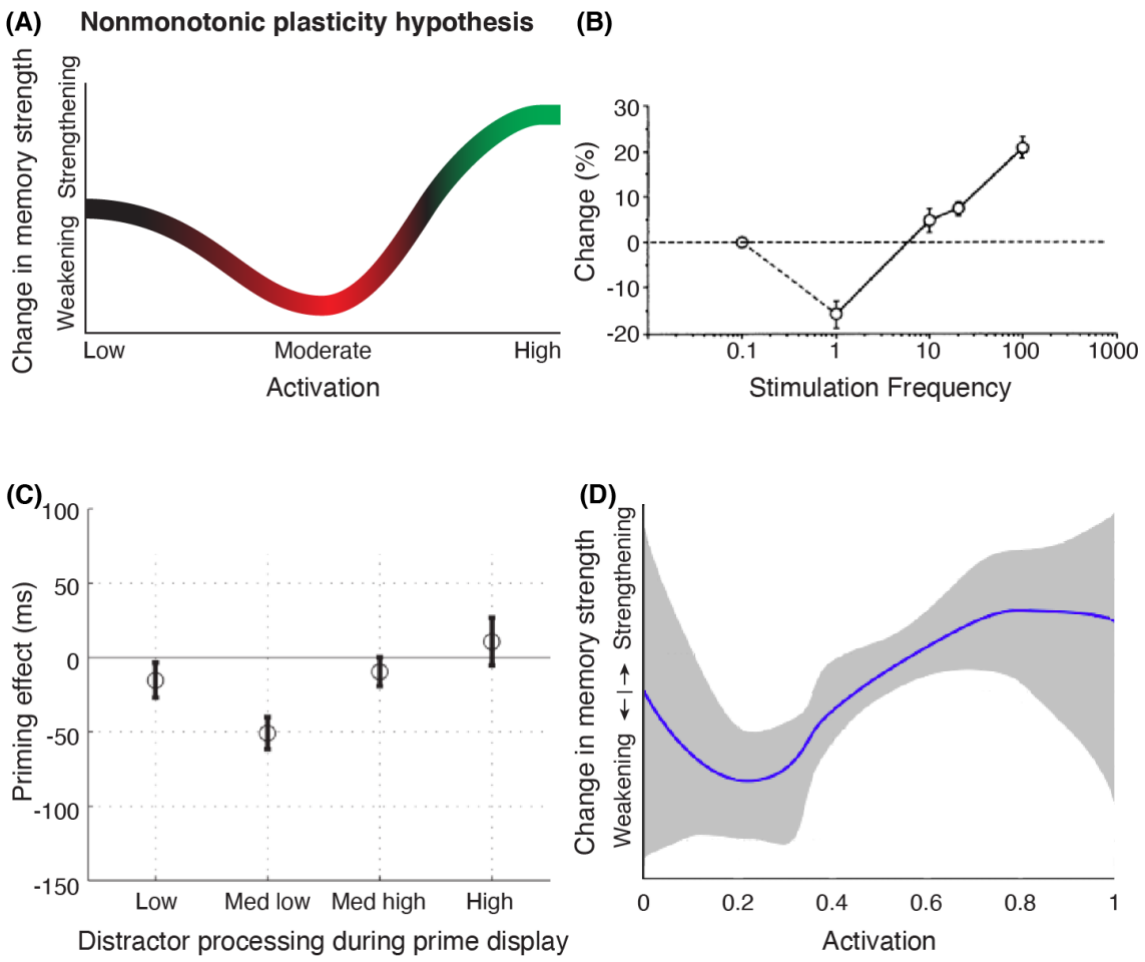


Figure 1.3: The Nonmonotonic Plasticity Hypothesis. (A) *The basic form of the nonmonotonic plasticity hypothesis (NMPH) is shown: Low activation causes no learning, moderate activation causes weakening of synaptic connections, and higher levels of activation cause strengthening of synaptic connections. The most prominent computational instantiations of the NMPH (e.g., [46]) also incorporate metaplasticity, whereby the "transition point" between strengthening and weakening is adjusted as a function of the average activity of the neuron; high average activity makes it easier to weaken connections into the neuron, and low average activity makes it easier to strengthen connections. This metaplasticity principle is well supported by neural data [53]; functionally, it is important for preventing runaway synaptic modification (whereby strong connections keep getting stronger).* ***(B)*** *U-shaped plasticity in rats is shown (adapted from [88]). Neurons in brain slices from rat visual cortex were stimulated at varying frequencies, then the slices were analyzed for long-term depression or long-term potentiation. Lower frequency stimulation produced long-term depression whereas higher frequency stimulation produced long-term potentiation.* ***(C)*** *There is also evidence of the U-shaped function in studies with humans using EEG and fMRI, showing that the NMPH scales beyond the synapse to account for human behavior. For example, these effects were seen in a negative priming study [50] where participants had to ignore perceptual distractors and then respond to them later. In this study, to-be-attended targets and to-be-ignored distractors came from different categories (e.g., face, scene), making it possible to track processing of the distractor using category-specific EEG pattern classifiers. Moderate levels of distractor processing were associated with negative priming (slower subsequent responding to the stimulus), but higher and lower levels of distractor processing were not associated with negative priming. Adapted from [50].* ***(D)*** *As another example, a study [49] obtained evidence for the NMPH in paired-associate learning, using a variant of the think/no-think paradigm [89]. Participants learned paired associates; later, they were given the cue (the first item in the pair) and asked not to think of the associated item. Moderate activation of the "no-think" associate during the no-think trial (measured using an fMRI pattern classifier) led to diminished performance on a subsequent memory test, whereas higher activation led to enhanced performance on a subsequent memory test (adapted from [49]; the gray ribbon in the figure indicates the 90% credible interval, meaning that 90% of the probability mass of the curve distribution is contained within the interval). Note that several other fMRI studies using different paradigms have obtained similar U-shaped results [54–56].*

when memories are highly activated—most prominently, the BCM learning rule [46,47], but also others [1,48]. We refer to the "U-shaped" learning function shared by all of these rules as the *nonmonotonic plasticity hypothesis* (*NMPH*; [49,50]). The NMPH is supported by neurophysiological data showing that moderate postsynaptic depolarization leads to long-term depression (synaptic weakening) and stronger depolarization leads to long-term potentiation (synaptic strengthening; [51–53]). Furthermore, recent studies in humans show that the NMPH "scales up" to explain human behavioral and neural data: Moderate activation of memories (measured using multivariate fMRI or EEG) is associated with subsequent forgetting of those memories and higher activation is associated with subsequent improvements in memory (e.g., [49,50,54–56] **(Fig 1.3)**.

A useful way to think of the NMPH in relation to supervised learning is to focus on where the learning "targets" come from: Supervised learning adapts synaptic weights so activation patterns in the network match explicit targets supplied directly by the environment. With the NMPH, the spread of activation throughout the network sets the targets: Weights are adapted to reinforce strongly activated patterns and to destabilize moderately activated patterns, at all layers of the network. In this framework, strongly activated memories (e.g., retrieving "hollow" from the cue "absence") can be viewed as internally generated targets for learning.

How the NMPH accounts for representational change

In addition to explaining how memories get stronger and weaker, the NMPH also makes detailed predictions about representational change [1,57]. When one memory is strongly activated and an overlapping memory is moderately activated at the same time, the NMPH predicts that connections from the strongly active memory to the moderately activated memory

will be weakened, leading to differentiation; but when the overlapping memory is strongly activated, connections will be strengthened, leading to integration [57] **(Fig 1.4)**. The functional significance of these changes is to reduce competition on subsequent retrieval attempts. This

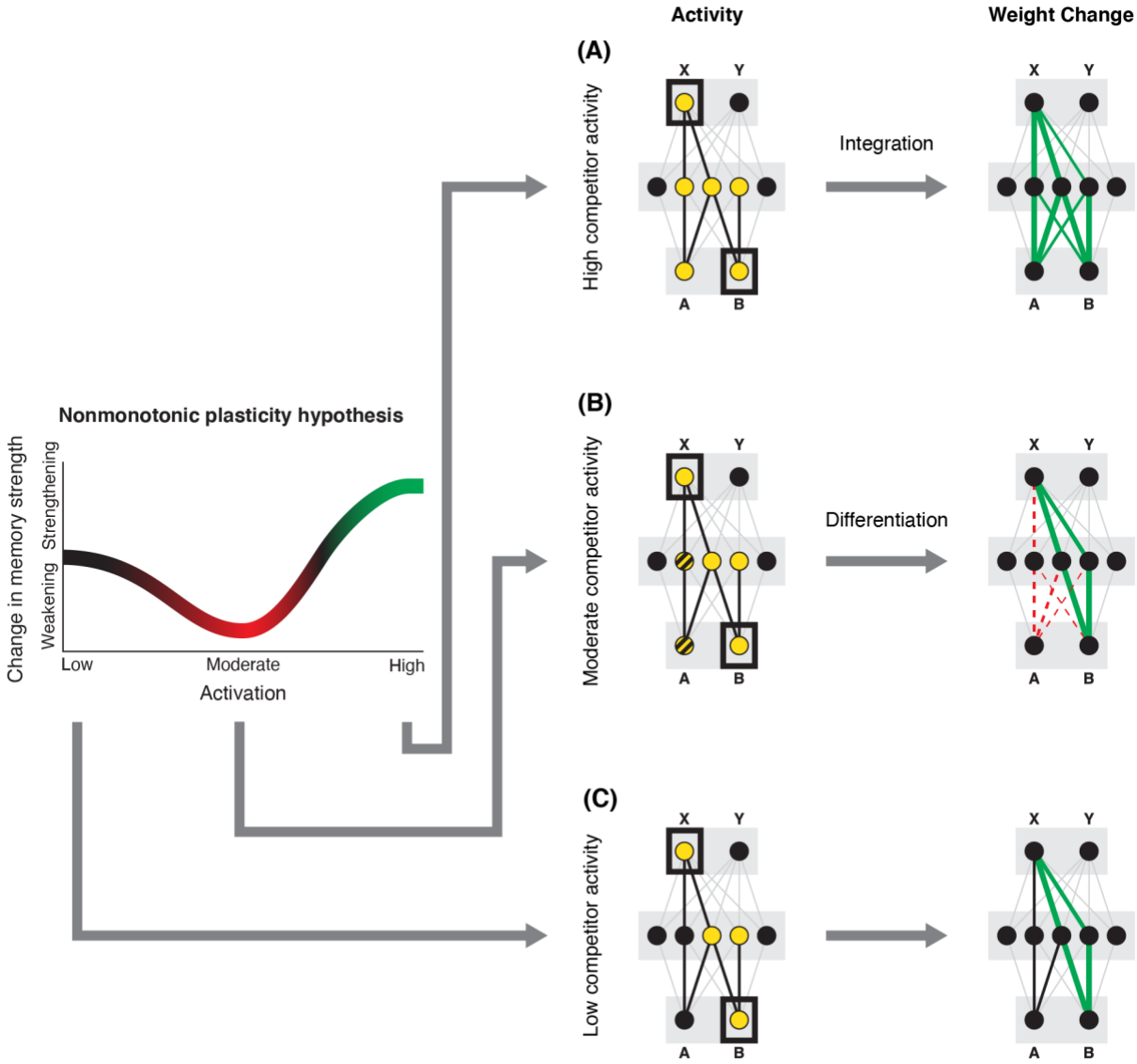


Figure 1.4: How the NMPH leads to differentiation and integration. *This figure depicts the same situation as the supervised learning figure (**Fig 1.1**). A has been linked to X, and B has some initial hidden layer overlap with A. In this network, activation is allowed to spread bidirectionally. When B is presented along with X (corresponding to a B–X study trial), activation can spread downward from X to the hidden layer units associated with A, and also—from there—to the input-layer representation of A. **(A)** If activation spreads strongly to the input and hidden representations of A, integration of A and B occurs due to strengthening of connections between all of the strongly-activated features (green connections indicate strengthened weights; A–B integration can be seen by noting the increase in the number of hidden units receiving projections from both A and B). **(B)** If activation spreads only moderately to the input and hidden representations of A, differentiation of A and B occurs due to weakening of connections between the moderately activated features of A and the strongly activated features of B (green and red connections indicate weights that are strengthened and weakened, respectively; A–B differentiation can be seen by noting the decrease in the number of hidden units receiving strong connections from both A and B—in particular, the middle hidden unit no longer receives a strong connection from A). **(C)** If activation does not spread to the features of A, then neither integration nor differentiation occurs.*

happens regardless of whether learning manifests as differentiation or integration—both reduce competition, by transforming a situation with two competing memories to either a situation with one memory (integration) or two memories that are farther apart (differentiation). Intuitively, this corresponds to two ways of preventing kids from fighting—you can make them friends (integration) or you can separate them (differentiation) [58].

This reduction in competition is also how the NMPH accounts for the testing effect: During retrieval, activation spreads to related memories, causing the representational changes outlined above, which reduce competition on subsequent retrieval attempts and thus improve accuracy [34]. In contrast, during restudy—which improves memory less than retrieval—activation is largely limited to the restudied memory itself. Intuitively, activation spreads less far during restudy compared to retrieval because the (highly active) restudied item laterally inhibits other memories (for a neural network model of this phenomenon, see [59]). Because activation does not spread as far, there is less representational change and thus less impact on accuracy. A key prediction of this account (yet to be tested) is that the size of behavioral testing effects should correlate with the amount of representational change (e.g., measured with fMRI).

According to the NMPH, when a competing memory comes to mind moderately, it is both differentiated from the more-strongly-activated target memory and also weakened overall, making it harder to retrieve (insofar as the elements of this memory are now less-strongly interconnected and provide less mutual support to each other during retrieval). This explains the phenomenon of *retrieval-induced forgetting* (*RIF*), whereby retrieving one memory impairs the subsequent retrievability of related memories [59–62].

Furthermore, if the competing memory is subsequently restudied (or successfully retrieved, despite having been weakened), the NMPH posits that the memory will be re-

strengthened (due to strong coactivity of the constituent parts of the memory), while staying differentiated from the target. A key implication of this view is that the damaged-then-repaired memory is better off than before the damage occurred, because it has now been differentiated from the target memory and will suffer less competition from it. This leads to the prediction that memories that compete and are then restudied should be better remembered than memories that are merely restudied without having competed. This is what was observed by [63] and [57], who termed this phenomenon *reverse RIF*. Moreover, the size of this reversed behavioral effect can be predicted by the degree of neural differentiation of representations of the competing stimuli in the hippocampus [57]. Another prediction arising from the NMPH is that, in a RIF (or reverse RIF) paradigm, strong (as opposed to moderate) activation of a competing memory during retrieval should lead to integration of that memory with the retrieved memory, boosting its later accessibility. This has not been tested with fMRI, but behavioral evidence fits this prediction (e.g., [64]); see also [65] for related fMRI evidence from an associative inference paradigm.)

Two recent studies of statistical learning [55,66] provide additional support for the NMPH. In these studies, participants viewed a stream of stimuli with embedded regularities (e.g., scene A was regularly followed by scene B), which were later violated (e.g., scene A was followed by face X instead of scene B). Moderate activation of the predicted scene (B) in the brain at the moment the prediction was violated led to impaired subsequent memory for B [55]. In a follow-up study [66], B was re-studied after the violation; the same circumstance that led to forgetting before (moderate activation of scene B when A was unexpectedly followed by X) was now associated with neural differentiation of the hippocampal representations of A and B. This fits with the idea that moderate activation of B that co-occurs with strong activation of A leads to "shearing" of B's connections with A (for an alternative explanation, see [67]; for a related

result, see "weak pairs" in [68]). Note that these statistical learning findings could also potentially be explained by supervised learning (i.e., weakening of an incorrect prediction). The place where NMPH learning diverges from supervised learning is when presenting a cue triggers very strong anticipatory activation of the predicted item and the prediction is violated. In this case, the NMPH posits that integration of the cue and predicted item will occur, boosting the association, whereas supervised learning predicts that the association will be weakened (due to the large prediction error). In cases like this (where supervised learning and the NMPH make opposite predictions), the overall learning outcome will depend on the balance between supervised and unsupervised learning (see "Balancing Supervised and Unsupervised Learning").

Crucially, the NMPH can also explain the challenging data discussed earlier from [10,11], where linking two items to a common associate caused differentiation in some brain regions (e.g., the hippocampus; see also [12,31]). In this situation, linking the items to a shared associate provides an additional conduit for activation to spread from one item to the other. If the baseline level of spreading activation was low, linking to a shared associate may push spreading activation to moderate levels, which will (according to the NMPH) lead to differentiation. However, a key implication of the NMPH is that there are limits on this dynamic; if the other item activates strongly (instead of moderately), the NMPH predicts integration will result—potentially explaining the integration effects observed in some brain regions by [11] **(Fig 1.5)**.

This account highlights a useful property of the NMPH: It can explain qualitative differences between brain areas (in whether they tend to show integration or differentiation) in terms of a continuous, quantitative difference in how strongly competitors are allowed to come to mind. Specifically, brain areas differ in terms of how tightly excitatory activity is controlled by inhibitory interneurons (e.g., [69]). According to the NMPH, if it is relatively hard for

competitors to come to mind in a particular area (because of high inhibition), that area will favor differentiation; if it is easier for competitors to come to mind (because of lower inhibition), that area will favor integration. The fact that many of the differentiation effects reported in this paper involve the hippocampus may be a consequence of high inhibition and thus sparse activity in the hippocampus [28,69], which makes it difficult to strongly activate competitors. This view also

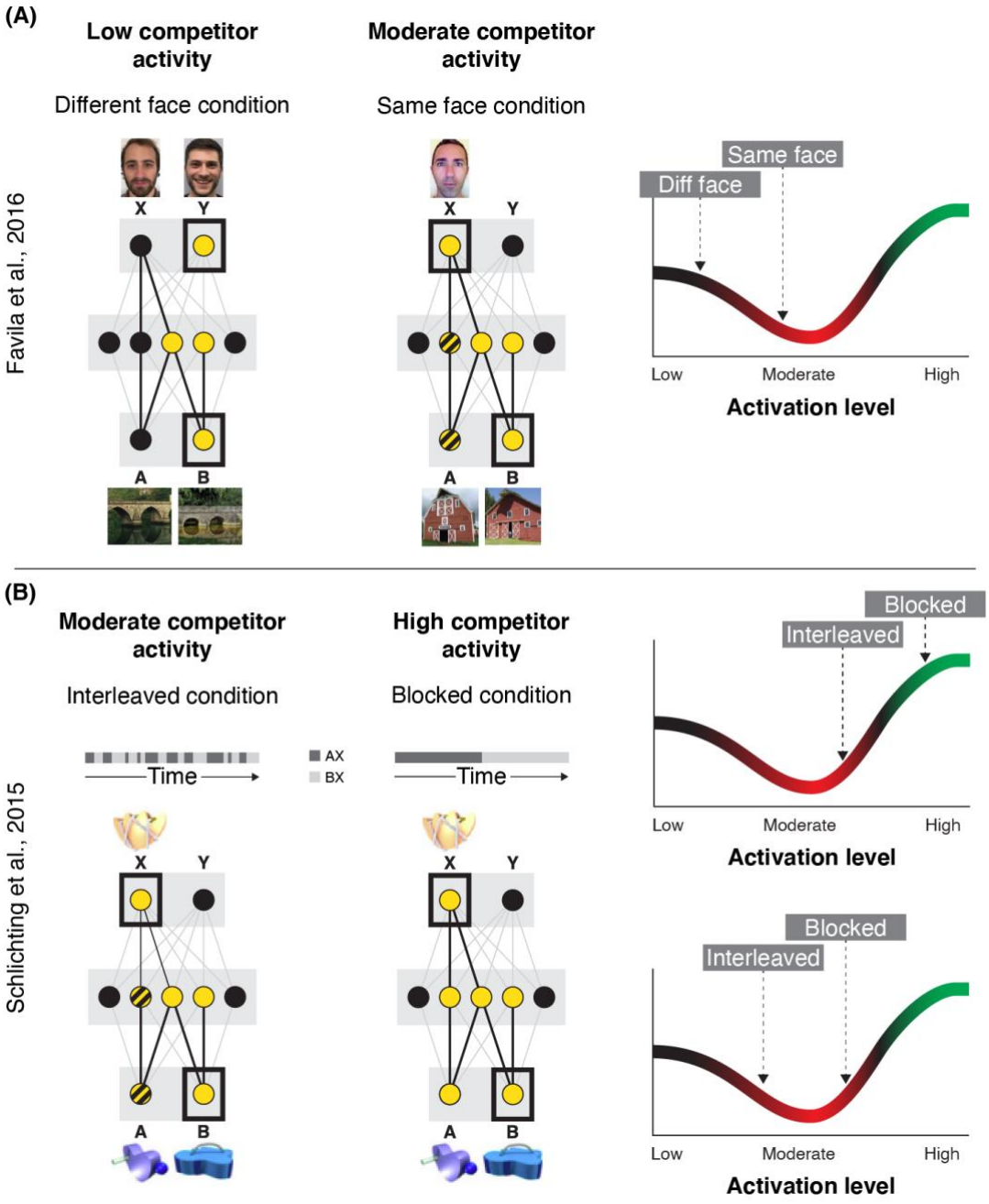


Figure 1.5: How the NMPH explains findings that challenge supervised learning. (A) *The NMPH explanation for the hippocampal differentiation results from Favila et al. [10] are shown. The network diagrams show the state of the network after scene A has been associated with face X, during a trial in which scene B is being linked either to face Y (different face condition) or face X (same face condition). In the different face condition, we assume that the level of activation spreading from scene B to its pairmate (A) in the hippocampus is relatively low. In the same face condition, the strengthened connection between X and A provides a conduit for activation to spread back down to A, resulting in A's representation becoming moderately active; this leads to differentiation of A and B, via the mechanisms shown in* ***Fig 1.4****. **(B)** To account for the results from Schlichting et al. [11], we use the same logic. The only extra dimension is how to account for the blocked versus interleaved training manipulation. In the blocked condition, all A–X learning occurs before any B–X trials take place; as such, the association between X and A's hidden representation will be stronger in the blocked than the interleaved condition. This strengthened A–X connection allows more activation to spread to the representation of A (in the input and hidden layers) during B–X trials. As noted earlier, a key feature of the results from [11] is variance across brain regions as to whether integration or differentiation was observed. The NMPH can explain between-area differences in integration/differentiation in terms of between-area differences in how strongly activation spreads to A during B–X trials. The network diagrams and the topmost NMPH diagram in B depict a situation where activation spreads moderately in the interleaved condition and strongly in the blocked condition; in this situation, we would expect differentiation in the interleaved condition and integration in the blocked condition. Alternatively, if overall levels of activation are lower in a given area, one might observe differentiation in both conditions (lower NMPH diagram), and if levels of activation are higher, one might observe integration in both conditions (not shown). This account also predicts that some configurations will not be observed; in particular, if activation is higher in blocked than interleaved, there is no way to position these conditions along the NMPH curve such that blocked leads to differentiation and interleaved leads to integration. It is therefore notable that, across the three main regions of interest surveyed by [11] (hippocampus, mPFC, and IFG), all three of the "possible" configurations were observed but the "impossible" configuration was not observed (see* ***Fig 1.2****).*

fits with the aforementioned finding [12] that—within the hippocampus—subregions with very sparse activity (CA2/3/DG) [69] tend to show differentiation, whereas subregions with less sparse activity (CA1) [69] tend to show integration, as well as the finding from [11] that posterior hippocampus (hypothesized to use sparser neural codes than anterior hippocampus [70]) tends to show differentiation, whereas anterior hippocampus tends to show integration. Another factor that may contribute to differentiation in the hippocampus is that it has a high learning rate, compared to cortical regions [2]. Because of this high learning rate, a single instance of competition may be enough to repel similar items a large distance apart in representational space, past the "baseline" level of overlap for dissimilar items.

As noted by [11] (see also work on complementary learning systems by [2]), having different areas be differentially biased toward integration vs. differentiation confers major functional benefits. If all brain regions exhibited the same level of bias toward integration or differentiation, the brain would have to choose whether to (globally) integrate or differentiate stimuli. Having areas with different biases toward integration vs. differentiation (due to different

levels of local inhibition) makes it possible to avoid this tradeoff. When multiple items are linked to a shared associate, some brain regions can integrate the items to promote generalization, and other regions can differentiate the items to ensure discrimination is possible.

Testing the NMPH

Going forward, how do we most effectively test the NMPH? A potential criticism is that—because the NMPH is flexible enough to explain both differentiation and integration—it is not falsifiable. Consider a pair of experimental conditions (call them X and Y) where memory activation is higher in condition Y than X. If representations in condition Y show less overlap (on average) than in condition X, one could explain this *post-hoc* by arguing that X and Y fall on the descending part of the U-shaped curve (where more activation leads to weakening and differentiation). Likewise, if representations in condition Y show more overlap than in condition X, one could explain this *post-hoc* by arguing that X and Y lie on the ascending part of the U-shaped curve (where more activation leads to strengthening and integration). Indeed, this kind of post-hoc theorizing can be found in **Fig 1.5**. This would be less of a concern if there were an *a priori* way of knowing whether conditions would elicit low, moderate, or high activation (thereby "locking down" whether those conditions should lead to differentiation or integration), but in practice this is hard to do.

As discussed in **Fig 1.5**, the NMPH cannot explain every possible outcome; however, it still is compatible with a wide range of findings, making it challenging to test. The key to robustly testing the NMPH account of representational change is to obtain samples from the full "x-axis" of the U-shaped curve, thereby making it possible to reconstruct the full U-shaped pattern of plasticity predicted by the NMPH. Crucially, some results have already been obtained

showing a U-shaped relationship between memory activation and subsequent memory behavior (e.g., [49]; see **Fig 1.3**), providing support for the NMPH. However, a similar result has not yet been obtained in studies of representational change (showing a smooth transition from no representational change to differentiation to integration within a single study)—this is a goal of our ongoing research.

Combining adaptive design optimization methods [71] (where the design of an experiment is changed on-line) and real-time fMRI [72] may also be useful for testing the NMPH. This approach would involve tracking memory activation during fMRI scanning using real-time decoding [72,73], keeping track of which points on the activation continuum have been sampled, and adaptively changing the task to better sample regions of the continuum that have been undersampled. For example, [74] showed that it is possible to parametrically nudge memory activation levels by varying concurrent distraction during memory retrieval using a multiple-object-tracking task. This suggests an approach whereby distraction could be increased to better sample lower regions of the activation continuum if memory activation is too high, and decreased to better sample higher regions of the activation continuum if memory activation is too low. If eliminating distraction during memory retrieval is not sufficient to elicit strong activation of the memory, another option is to "fade in" the to-be-retrieved stimulus onscreen during retrieval (for a related approach see [75]).

What is the function of the NMPH?

Our main claim thus far has been that supplementing supervised learning with unsupervised NMPH learning may help explain data on representational change and retrieval-driven learning. Here, we revisit the question of why the brain would incorporate NMPH

learning. Put another way, what functional benefits does the NMPH confer, when added to supervised learning?

We hypothesize that the need for NMPH learning derives from the problem of competition in recurrent neural networks, where there are reciprocal interactions between layers. These networks can exhibit complex settling dynamics before reaching a stable state; if the network vacillates between nearby states without fully settling in one (as can easily happen in neural network models with noise), this can prevent the system from acting decisively in response to the current stimulus, with potentially catastrophic consequences. The NMPH could help with this problem: its primary functional consequence is to reduce competition on subsequent retrieval attempts, which should lead to a faster and more accurate response (because of less vacillation, and a greater likelihood of eventually settling into the correct state).

At the same time, unsupervised learning has its costs—it can cause networks to become entrenched in particular knowledge states and make them less able to adapt to new inputs [43]; also, in pathological cases, it can lead the network to erroneous conclusions—with the NMPH, strongly retrieving a false memory can further embed that incorrect state. Our conjecture is that competition is a large enough problem to justify the need for an extra "housekeeping" process (implemented by the NMPH) that grooms the attractor landscape to reduce competition. Specifically, we hypothesize that the potential costs of unsupervised NMPH learning (e.g., relating to entrenchment) are outweighed by the benefits of reduced competition.

More simulation work is needed to assess whether this hypothesis is true. This question can be addressed by adapting recurrent neural networks that do not presently incorporate NMPH learning (e.g., the visual object recognition model described in [76]), and testing whether they benefit from the addition of the NMPH; for preliminary evidence that this is the case, see [77].

These ideas build on other work describing synergies between supervised and unsupervised learning [39,78–80], including a recent model of hippocampal learning [81].

Balancing supervised and unsupervised learning

One final, important question is how the brain sets the balance between supervised and unsupervised learning. As noted above, this question is especially relevant when considering situations where the NMPH and supervised learning make conflicting predictions. Speculatively, in light of the idea that NMPH learning serves to reduce competition on future retrieval attempts, it might be useful to ramp up the influence of NMPH learning (relative to supervised learning) when competition is persistently high. Accomplishing this requires two things: a means of adjusting the influence of NMPH learning, and a control mechanism that can detect competition and trigger changes in the influence of NMPH learning.

With regard to the first issue (how to adjust the contributions of NMPH), one intriguing idea is that these contributions could be titrated by varying local levels of cortical excitability, e.g., by varying the amplitude of inhibitory oscillations [1,59]. Lowering the local level of inhibition allows competing memories that are presently inactive (because inhibition outweighs excitation) to become active; this has the effect of pulling these memories from the far-left side of the U-shaped curve to the moderate-activation region associated with differentiation or even the high activation region associated with integration. Thus, by increasing the extent to which competing memories come to mind, larger-amplitude oscillations could lead to more representational change, which would help mitigate competition and boost subsequent memory. Although existing studies have not directly tested this claim, it is consistent with data showing that theta oscillation amplitude at encoding predicts subsequent retrieval success (e.g., [82–84]).

If the brain up-regulates NMPH learning by up-regulating inhibitory oscillations, what triggers that change? Under the assumption that NMPH learning should be ramped up when competition is high, it would make sense to have brain regions that are involved in detecting and resolving retrieval competition (e.g., anterior cingulate cortex and ventrolateral prefrontal cortex [85]) control the strength of oscillations. Indeed, a possible path for this exists: Ventral prefrontal regions project to the basal forebrain, which controls the release of acetylcholine, which (in turn) is known to modulate oscillation strength (for a review, see [86]).

Concluding remarks and future directions

Determining when learning leads to differentiation vs. integration has been a challenging undertaking for memory researchers. It is hard to imagine a more fundamental question, and the fMRI data reported here show that it is a complex business, full of seemingly contradictory experimental results. We have argued that supplementing supervised learning with nonmonotonic unsupervised learning helps to explain these divergent findings and a wide range of other learning phenomena, including the testing effect, (reverse) retrieval-induced forgetting, and effects of expectation-violation in statistical learning. Although there is still much more work to be done, the research reviewed here shows that the flow of activation through neural networks can powerfully sculpt memories that are activated, strengthening or weakening them, and pushing them together or pulling them apart, in ways that cannot always be explained by supervised learning. Now that we have the means to track both memory activation and the resulting representational change in humans with fMRI, we expect rapid progress in characterizing these unsupervised learning effects in the years to come.

Chapter 2: Studies of how competition affects memory for continuous stimulus values

Introduction

Chapter 1 discussed how the NMPH could explain the differentiation of similar memories seen in several fMRI studies [10–12]. We also discussed how the NMPH could explain changes to the retrievability of memories. In this chapter, we investigated how the NMPH impacts not only the retrievability, but also the *contents* of memories in a series of behavioral experiments.

Recall the retrieval induced forgetting (RIF) paradigm briefly described in Chapter 1. The typical RIF paradigm [60–62] involves an initial study phase, in which participants learn category-exemplar pairs (e.g., "Fruit-Orange," "Fruit-Banana," "Metal-Chrome," "Metal-Platinum"). Then, during the retrieval practice phase, half of the items in half of the categories are cued for retrieval (e.g., "Fruit-Or___"); these are called Rp+ items. After multiple rounds of retrieval practice for Rp+ items, all items undergo a final cued recall test. Typically, the Rp+ items (here, "Fruit-Orange") show improved recall in comparison to unpracticed baseline categories ("Nrp;" here, any of the "Metal" word pairs). Crucially, memory for the items that were not practiced but were categorically related to words that were practiced ("Rp-;" here, "Fruit-Banana") is worse than memory for the baseline Nrp words. This effect is reversed ("reverse RIF", or "revRIF") in a paradigm where both Rp- and Nrp items are allowed to be restudied. When this happens, the memory for Rp- items is better than the memory for Nrp items [57,63]. These results are puzzling because Rp- and Nrp are studied the exact same number of times, yet restudy of Rp- and Nrp leads to an increase in retrievability of Rp- items only.

As explained in Chapter 1, the NMPH could explain these findings, and predicts that the relatively better recall of Rp- item arises from differentiation between Rp- and Rp+ items. Retrieving "Fruit-Orange" will moderately activate other Fruit item competitors (including "Fruit-Banana"); this will weaken the Banana representation as well as the connections between Orange and Banana, initiating the process of neural differentiation. However, restudying "Fruit-Banana" will re-strengthen the Banana representation. The net result of Orange and Banana competing during retrieval of "Fruit-Orange," followed by restudy of Banana, is strong and differentiated neural representations. Hulbert & Norman (2015) [57] provided evidence for the NMPH's role in these effects, showing reduced pattern similarity between memory competitors was correlated with retrieval-induced facilitation of Rp- items in the revRIF paradigm.

However, there is another potential prediction of the NMPH for revRIF—the shift in the contents of memory. Just as the weakening and strengthening of neural connections lead to changes in how easily the Rp+, Rp-, and Nrp items are remembered, representational change between Rp+ and Rp- as shown in Hulbert & Norman (2015) [57] should also lead to distortions in how these memories are recalled. The NMPH predicts Rp- is restored in the revRIF paradigm through the addition of new neurons (during restudy) that were not previously associated with either the Rp+ or Rp- item. An outcome of this might be that the Rp- item is remembered in a distorted way, biased away from details of the memory shared with Rp+ (**Fig 2.1**).

The goal of this research was to investigate the how the NMPH may impact behavioral measures of memory for competing items. Critically, we wanted to separately measure the effects of retrieval practice on a competing item's *retrievability* (i.e., how easily it is remembered), and its memory *contents* (i.e., if it is remembered accurately or distorted). We developed a paradigm using a continuous feature space (i.e., color), to study this both within a

revRIF paradigm and in other paradigms. We used color because it is a continuous feature space making distortions easier to measure, and there is evidence that neurons in visual cortex are responsive to specific colors [90,91]. Moreover, there is existing evidence that color is susceptible to the effects of RIF [92], and consequently should also be susceptible to the effects of revRIF. We predicted that retrieval practice of an item (Rp+) should cause its pairmate (Rp-)

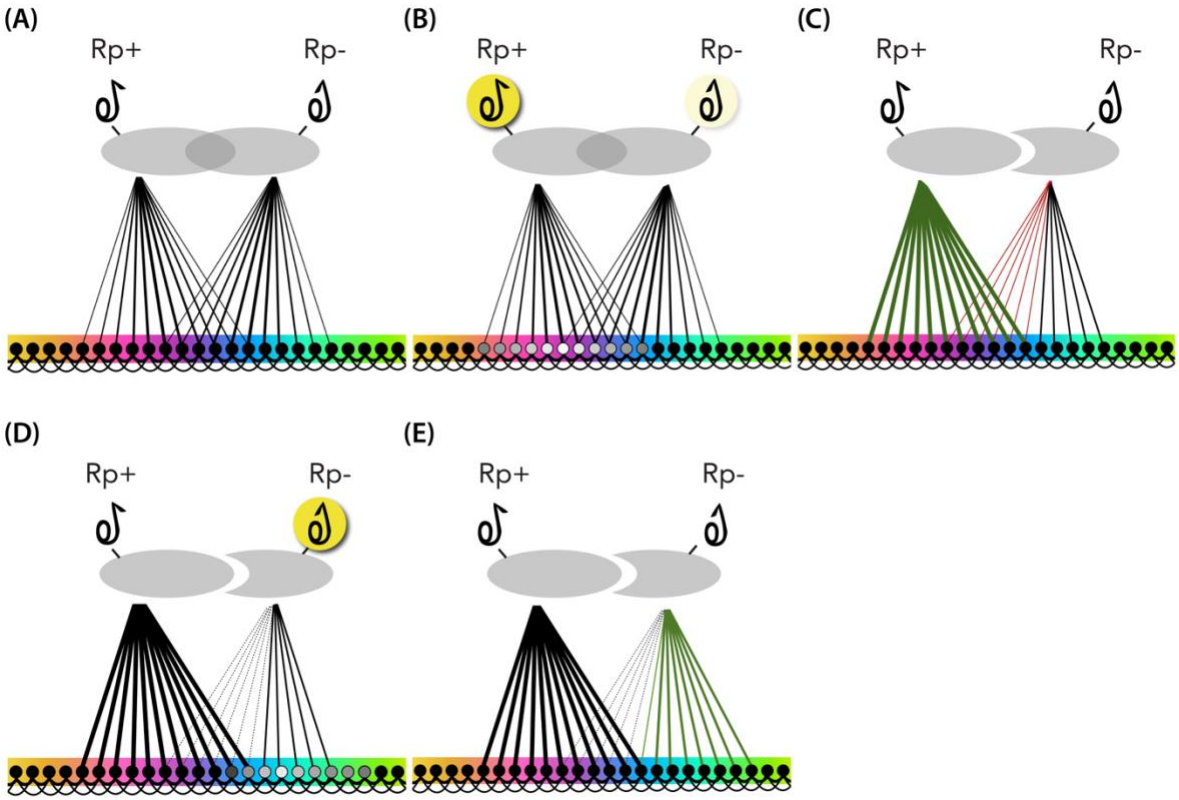


Figure 2.1: Theoretical model of revRIF's impact on color. ***(A)*** *Two partially overlapping memories are shown, with two similar colors. The top circles indicate the hidden layer representation for the Rp+ and Rp- shapes. The hidden layer has graded connections to a color layer, where the units are selective for specific colors and have recurrent connections to neighboring units. The connections from the hidden layer to the color layer are graded so that each item is most connected to a particular color unit and less connected to neighboring color units. There would be other, weaker hidden-to-color connections to further away color units (not shown). Inhibitory interneurons (not shown) enforce an approximate "set point" on the amount of neural activity in the network.* ***(B)*** *During retrieval practice of the Rp+ item, the color units with the strongest connections are activated most (i.e., purple), and further away color units are activated less (i.e., purple-red and purple-blue). This partial activation of blue-purple leads to partial activation of the Rp- shape.* ***(C)*** *Since the Rp+ shape units are highly active, the coactivity for connections to active colors are high, which triggers strengthening (highlighted in green). Crucially, since the Rp- shape units interfere and became weakly active, the connections between Rp- and the previously shared color units (purple-blue) weaken (highlighted in red).* ***(D)*** *When the Rp- item is restudied, previously connected color units (purple-blue) are no longer active. Other color units that were not previously activated take their place via spreading activation, meaning more blue-green units become active.* ***(E)*** *Because of the high coactivity between Rp- and these new units, Rp- regains connections to the color layer, but these connections are shifted away from the Rp+ color. This boosts retrievability of the Rp- color, but the recall color is distorted toward green.*

to move away, which would manifest behaviorally as color memory for the Rp- item biased away from the color of its Rp+ pairmate.

To preview, our studies did not show any conclusive evidence regarding distortions resulting from learning of similar memories; none of the studies we ran showed a color distortion effect away from competitors. One study did show a trend toward the predicted effect on retrievability (indicating that colors memories may be susceptible to the effects of revRIF). However, most studies we ran using revRIF did not show a retrievability effect, indicating that our procedure did not adequately put pairmates into competition with each other. We found it was difficult to create a task that is easy enough for participants to provide precise memory reports, while challenging enough to induce adequate competition. We outline the studies in this chapter so future researchers can see the limitations on these effects, and we distill several lessons which may expedite follow-up work in this field.

Experiment 1

We first adapted the revRIF paradigm from Hulbert & Norman [57] for color memories. We predicted effects both on the retrievability and distorted contents of memory. Specifically, we first predicted memory for both Rp+ and Rp- items would be better than memory for Nrp items. Assuming we found this effect on retrievability of the memories, we also predicted that memory for Rp- items would be distorted away from Rp+ items in color space.

Methods

24 naïve individuals (15 women, 8 men; mean age: 21 years old) participated in this experiment. The study protocol was approved by the Princeton IRB. The experiment took place

over two consecutive days. The first day lasted approximately 30 minutes and the second day lasted approximately 1 hour and 45 minutes.

Stimuli

Novel shapes were used, created freehand in a vector graphics drawing program (**Fig 2.2A**). Shapes were based on letters from extinct alphabets, sourced from an online repository of non-Roman ideograms (omniglot.com). We used 24 pairs of similar shapes (48 shapes total). Shapes within a pair were visually similar, with one feature rotated, reflected, or moved to a different location. Altering shapes in this way created a natural similarity between the shapes within a pair, while allowing them to be obviously distinct.

Each shape was associated with a distinct color (from a perceptually uniform space: equiluminant CIE *L*a*b** color space centered at L=70, a = 20, b = 38; radius 60). Color values were randomly sampled from 180 circular coordinates, with the restriction that colors of pairmates had to have a degree difference of 64˚. The 24 pairs of shapes were divided into 12 Rp pairs and 12 Nrp pairs. There were thus 12 Rp+ shapes, 12 Rp- shapes, and 24 Nrp shapes.

Participants were provided with a cover story as part of the instructions. They were told that they would encounter two self-portraits of aliens from various alien species. The two aliens within a species (i.e., the two pairmates) would look very similar to each other with only small differences. Every alien would also have its own name. The first three letters of a name can be thought of as a family name, and would be shared by the two aliens within a species. The participants were tasked with learning about these individual aliens.

These 24 family names were all of the form Vowel-Consonant-Consonant or Consonant-Vowel-Consonant. They were pronounceable in English, but were not homophones with any

English word. The individual alien names were made by randomly assigning one of –o –i –u or –a at the end of the family name. Thus, there were 48 four-letter names, one for each shape.

Procedure

The procedure is outlined in **Fig 2.2**. On the first day, participants completed a Name Learning Task to learn the three-letter "family names" for each pair. This task had a study and recall phase, and it only ended once all the names had been correctly provided (**Fig 2.2B**). Then, participants completed a Same-Different Task, where they had to tell whether two side-by-side images were the same exact shape, or different shapes; different shape trials were comprised of the two similar-looking pairmates (**Fig 2.2C**). This task was primarily meant to help participants learn to discriminate between the shapes. Color information was not shown on Day 1.

On the second day, participants repeated the Name Learning Task (recall portion only) and the Same-Different Task. Then, they did a Study Task where they studied the color and four-letter name for each shape (**Fig 2.2D**). Participants were told that studying the color would be important for a later test. Then, they did the Rp+ Retrieval Practice Task interleaved with the Rp-/Nrp Restudy Task (**Fig 2.2E**). During Rp+ trials, the shape was shown in white along with its name, and participants recalled the color using a color wheel. The item in its correct color was then shown as feedback. During Rp- and Nrp trials, the same information was shown, except the order of the trial elements was reversed—first, the correctly colored item was shown, and then participants were immediately cued to recall the color just shown. This means that Rp+, Rp-, and Nrp items were shown the exact same number of times, but the Rp+ trials required memory retrieval whereas the Rp- and Nrp trials did not (this mirrors the approach taken by Hulbert &

Norman, 2015) [57]. Then, participants did a Final Recall Task to get a final measure of color memories for all shapes. Finally, participants completed a third Same-Different Task.

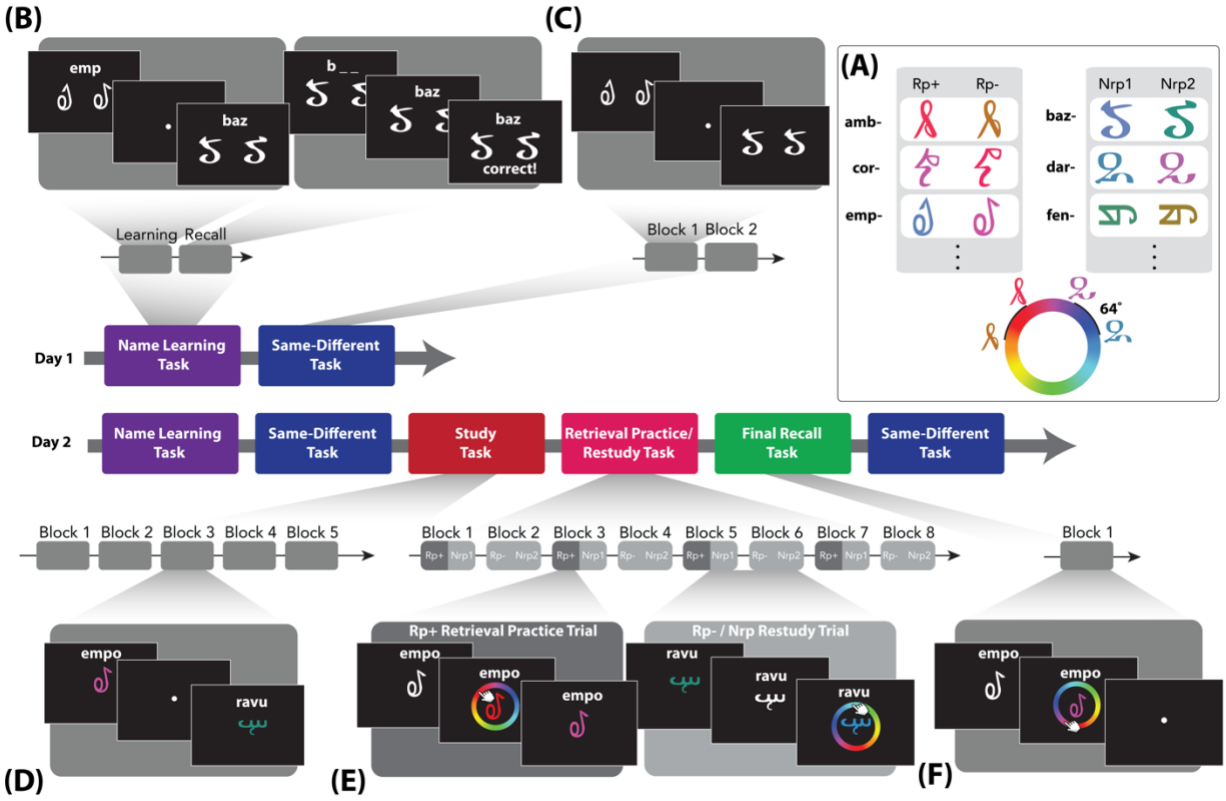


Figure 2.2: Procedure for Experiment 1. ***(A)*** *24 pairs of shapes were used (48 shapes total). The pairs were divided into 12 Rp pairs and 12 Nrp pairs. Each "alien" pair had an associated three-letter "family name," and each individual alien had a unique four-letter name starting with the family name. The colors of each shape were created by sampling the space around a color wheel, where the shapes within a pair had a color difference of exactly 64˚.* ***(B)*** *Name Learning Task. In the first part of the Name Learning Task, all family names were shown along with the two pairmates, in white. Participants studied each name and pressed a space bar to move on to the next name. Each pair was shown only once. In the recall phase, each pair was shown without the name. Participants were instructed to recall the name and type it in. If the name was correct, the display would say "Correct." Then, the pair was not shown again in the recall phase. If the entered response was incorrect, the display would show "Incorrect. The correct name was:" along with the correct family name. The pair would then be presented again during one of the subsequent trials until it was correctly named.* ***(C)*** *Same-Different Task. The task consisted of two blocks with 96 trials per block—48 "same" trials, and 48 "different" trials. Each of the 48 items was shown once in a "same" trial, and each pair was shown twice in "different" trials. Participants were shown the shapes without the associated name, and used the "z" or "/" keys to indicate whether the two items on screen were identical or different (i.e., pairmates). The shapes appeared for 1 s, and participants had a maximum of 5 s to make their response before the next trial started.* ***(D)*** *Color Study Task. In each of the five blocks, each of the 48 colored shape appeared once with its full, four-letter name for 5 seconds. Participants were instructed to study the name and color, and were encouraged to make a vivid mental picture of the alien who would draw this picture. A 2 s black screen followed, where participants used the keys 1,2,3, or 4 keys to rate how vivid their mental picture of the alien was.* ***(E)*** *Retrieval Practice/Restudy Task. The eight blocks alternated between Rp+/Nrp1 blocks, and Rp-/Nrp2 blocks. For the Rp+ trials, the shape appeared for 2 s in white along with its name. Then the color wheel appeared around the shape, and participants used the cursor to select the color they remembered; as they moved around the color wheel with the cursor, the color of the shape changed. They had a maximum of 30 s to make their response. After they made a response (or 30 s elapsed), feedback was shown, in the form of the correct colored shape for 2 s. For the Rp- and Nrp restudy trials, the stimulus presentation and timing were identical but reversed—the correct color was shown first, and then participants recreated the color just presented. The color wheel was rotated for each new trial.* ***(F)*** *Participants completed a Final Recall task which consisted of one block where each of the 48 items were shown once, in white. Participants used the color wheel to make their final memory report for each. The shape appeared in white for 2 s and then participants had a maximum of 30 seconds to make their response.*

Color Reports

Responses on the Final Recall Task were calculated as the angular deviation from the position on the wheel corresponding to the correct color value. For example, if the correct value was at 30˚, a response of 40˚ was coded as either +10˚ or -10˚. The sign depends on whether the color was toward (+) or away from (-) the competitor's color.

Mixture Model

For each condition, we fitted a mixture model to separate the parameters of interest [93,94]. Color errors could be the result of several factors, and thus can be modeled by a mixture of three distributions (**Fig 2.3A**): 1) a uniform distribution to reflect a random guess of any of the potential degree values; 2) a von Mises (circular normal) distribution around the correct value to reflect an accurate but imprecise correct memory representation; 3) a von Mises distribution around the competitor's color to reflect a swap error, where the pairmate is recalled. Maximum likelihood estimation was used to recover the likelihood that responses were drawn from each of these distributions by fitting parameters that characterize each of these distributions [93].

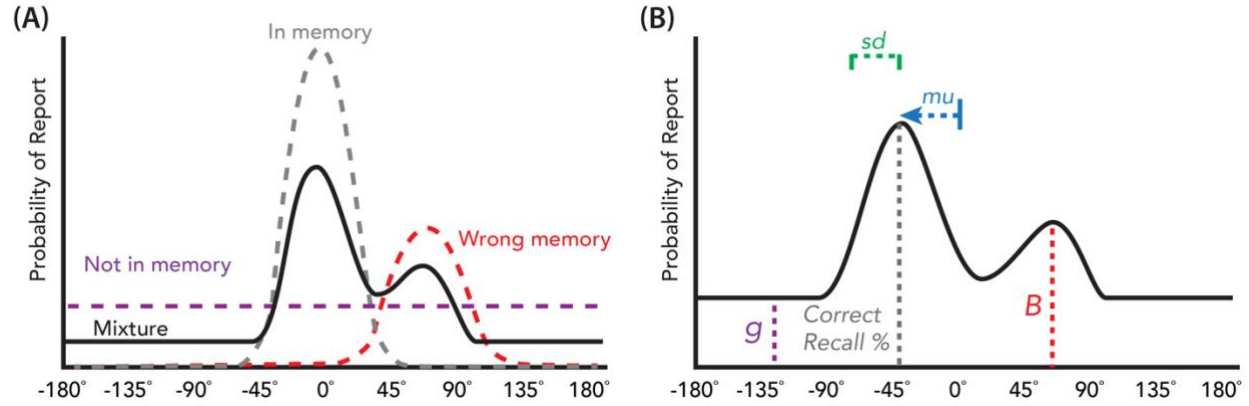


Figure 2.3: Mixture Model. (A) *The errors could be a result of three distributions: 1) When no color is present in memory and the participant makes a guess, the reported color is equally likely to be anywhere on the circle (purple). 2) When the report comes from the correct memory, the reported color is likely to be near the true color, 0˚ (grey). 3) When a swap error occurs and the report comes from a recollection of the pairmate's color, the reported color is likely to be near the degree for that color, here 64˚ (red). These distributions are collapsed in our mixture model (black).* ***(B)*** *The parameters of interest used in our model, as defined in the text. A negative mu indicates a shift of the correct recall distribution away from the pairmate's color (here, the shifted 45˚ away from the competitor).*

There are several parameters that were used in this model (**Fig 2.3B**): 1) the height of the uniform component, which represents the probability of guessing, *g*; 2) the standard deviation of the correct von Mises component, *sd*; 3) the bias, *B*, which represents the probability that the response was pulled from the incorrect, pairmate distribution; and 4) the shift of the correct von Mises distribution to the left or right of the true value, *mu*. A *mu* of -10˚ would mean that the participant's error for that stimulus was 10˚ away from the pairmate. We also fitted parameters for the precision and shift of the pairmate (swap) distribution, but since we are interested in distortion effects when the item is correctly recalled, these pairmate distribution values are not reported in our results.

The mixture model we used fits the probability of two of the three distributions—*g* is the uniform distribution, and *B* is the pairmate distribution. We can calculate the probability that the memory comes from the *correct recall* distribution by subtracting *g* and *B* from 1.

Analyzing the parameters of interest

Our first goal was to examine the retrievability of memory. In a revRIF paradigm, we would expect the best memory for the Rp+ items and the worst memory for the Nrp items. We expected this to be reflected in the *correct recall* probability, such that the Rp+ items would have the highest probability of *correct recall* and the Nrp items would have the lowest probability. It could also be shown through the precision estimate, *sd*—we would expect Rp+ to have the highest precision (lowest *sd)* and Nrp to have the lowest precision (highest *sd).*

Our second prediction relates to the *mu* parameter, which shows the distortion of the *correct recall* distribution toward or away from the competitor. We predicted that if revRIF occurs (i.e., if the retrievability result above is found for *correct recall* and/or *sd*), then there

should be a distortion of the Rp- item away from the Rp+ item, but no distortion in the Nrp items. This would be shown by a negative *mu* for Rp- items.

Fitting an appropriate mixture model requires more observations than we collect per subject in the Final Recall Task. To compensate for this, we made a "super subject" dataset where we collapsed all the color errors from all participants together, which provides enough data points to fit the model. In order to test the validity of these parameters, we used a bootstrapping technique where 1000 super subjects were generated by sampling participants with replacement, and a mixture model was fit for each iteration. This provided a population-level 95% confidence interval (CI) for each parameter in terms of the proportion of bootstrap samples. We tested for significance by comparing the 95% CIs of the bootstrap samples to zero.

Results and discussion

Results for the model fitting parameters are shown in **Fig 2.4**. We did not find the typical revRIF effects on retrievability—the difference between probability of *correct recall* between Rp+ (M: 64.06%, 95% CI: 49.80% – 73.73%), Rp- (M: 67.10%, 95% CI: 51.80% – 76.80%) and Nrp (M: 60.38%, 95% CI: 39.92% – 70.26%) was not significant.

We also did not find any revRIF effects on precision, since the *sd* for Rp+ (M: 13.64˚, 95% CI: 10.34˚ – 16.94˚), Rp- (M: 16.36˚, 95% CI: 11.81˚ – 20.91˚), and Nrp (M: 15.33˚, 95% CI: 10.44˚ – 20.21˚) were not significantly different from each other.

The parameter *mu* represents the shift in color space, where a negative *mu* indicates a shift away from the pairmate. None of the fitted *mu* values were significantly different from zero, nor were there any differences between groups, indicating that there were no distortion effects

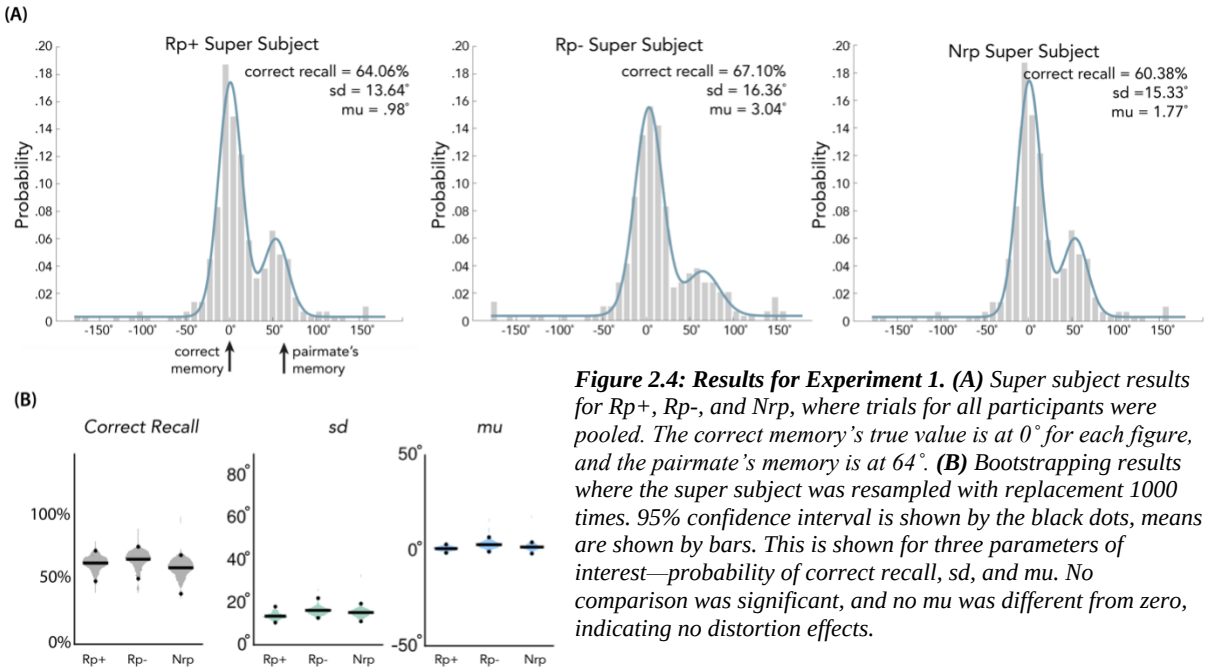


Figure 2.4: Results for Experiment 1. (A) *Super subject results for Rp+, Rp-, and Nrp, where trials for all participants were pooled. The correct memory's true value is at 0˚ for each figure, and the pairmate's memory is at 64˚.* ***(B)*** *Bootstrapping results where the super subject was resampled with replacement 1000 times. 95% confidence interval is shown by the black dots, means are shown by bars. This is shown for three parameters of interest—probability of correct recall, sd, and mu. No comparison was significant, and no mu was different from zero, indicating no distortion effects.*

(Rp+ M: .98˚, 95% CI: -1.10˚ – 3.05˚); Rp- M: 3.04˚, 95% CI: -.87˚ – 6.96˚; Nrp M: 1.77˚, 95% CI: -2.38˚ – 5.93˚).

Since we did not see the revRIF effects on retrievability, we would not expect any distortion to occur. In order for the Rp- item to be distorted away from Rp+, the Rp- item has to be degraded during the retrieval practice trials and then re-strengthened in the restudy trials. The fact that we did not see any effects on the *correct recall* or *sd* parameters indicates that this kind of competition-dependent learning did not occur in this study.

Experiment 2

In Experiment 2 we wanted to address why we may not have seen any revRIF effects in Experiment 1. One possible explanation is that the restudy of Rp- and Nrp items may have been too effective in Experiment 1. Seeing the correct color on Rp- trials could "re-center" the memory representation, preventing it from distorting away from the memory of the Rp+ item.

In this experiment, we wanted to provide some restudy in order to reactivate the Rp- memories, but not so much that the memory is anchored in place. We did this by providing restudy again for Rp- and Nrp like in Experiment 1, except this time restudy was just a brief flash of the correct color. We thought this could point participants in the right direction, but still leave their memory subject to distortion.

We also increased the degree difference between the colors of the shapes within a pair in order to prevent swap errors (*B)*, where participants incorrectly recall the color of the pairmate.

Methods

26 naïve individuals participated in this experiment (17 female, 9 male; mean age 20 years old). The procedure was the same as Experiment 1 (**Fig 2.2**), except for that we increased the difference between the two colors within a pair to 96˚ instead of 64, and we changed the duration of the restudy for Rp- and Nrp trials to be just 50ms instead of 2000 ms.

Results and discussion

Results for the model fitting parameters are shown in **Fig 2.5**. We did find evidence of an effect on retrievability of the memories, in that that the Rp+ items were most easily remembered and the Nrp items were least well remembered; the *correct recall* probability for Rp+ (M: 70.81%, 95% CI: 53.16% – 82.12%) was significantly higher than that for Nrp (M: 58.21%, 95% CI: 46.69% – 67.57%) ($p < .05$). The Rp- item correct recall probability (M: 65.48%, 95% CI: 51.49% – 77.79%) was numerically in between the two, and while the difference between Rp- and Nrp was trending ($p = .07$), it was not significant. These differences between the conditions indicate that reducing the quality of the restudy (by making it a 50ms flash) may have been

sufficient to create the revRIF effects, where both Rp+ and Rp- items are remembered better than baseline, Nrp items. While retrieval induced forgetting has previously been shown for color memory [92], this study is the first to suggest that color memory is susceptible to revRIF effects where memory for the Rp- item was boosted above baseline by interleaved restudy (although we should emphasize that the effect was not significant here). Additionally, the previous studies [92] used discrete colors that participants picked from a pool, and our result indicates that mixture models can be used to estimate retrievability from color on a continuous spectrum.

We also found a difference in precision (*sd*) between conditions; precision was best for Rp+ (M: 15.27˚, 95% CI: 11.36˚ – 18.06˚) and worst for Rp- (M: 24.44˚, 95% CI: 18.06˚ – 29.90˚) (p <.01). The difference between Rp+ and Nrp (M: 19.80˚, 95% CI: 17.25˚ – 22.53˚) was also significant (p <.01).

Since we found an effect on retrievability, we were particularly interested in distortion effects, represented by the parameter *mu*. However, none of the *mu* values for any of the

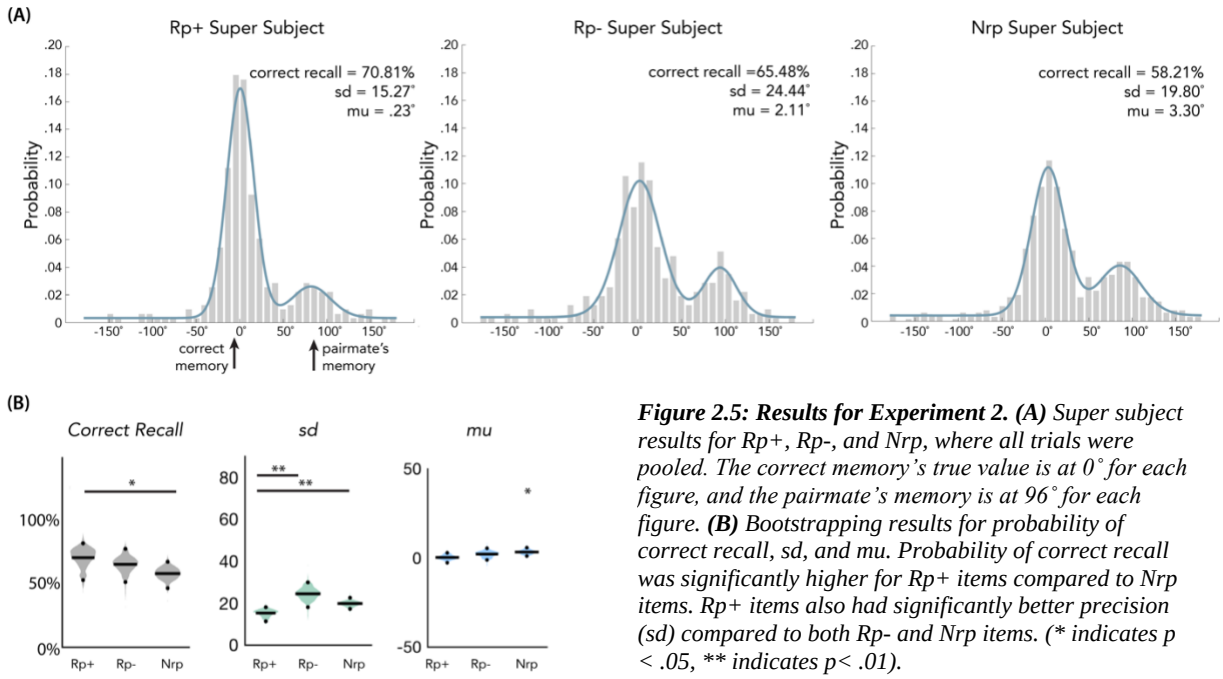


Figure 2.5: Results for Experiment 2. (A) *Super subject results for Rp+, Rp-, and Nrp, where all trials were pooled. The correct memory's true value is at 0˚ for each figure, and the pairmate's memory is at 96˚ for each figure.* ***(B)*** *Bootstrapping results for probability of correct recall, sd, and mu. Probability of correct recall was significantly higher for Rp+ items compared to Nrp items. Rp+ items also had significantly better precision (sd) compared to both Rp- and Nrp items. (* indicates p < .05, ** indicates p< .01).*

conditions were significantly different from each other, or less than zero (which would indicate distortion away from the competitor). The Nrp *mu* was positive (M: 3.30˚, 95% CI: 1.05˚ – 5.71˚, p <.01), indicating that the Nrp items were distorted a small amount toward the competitor. It is possible this happened because without any retrieval practice, the two Nrp items within a pair merge slightly. There were no distortion effects for the Rp+ (M: .23˚, 95% CI: -2.68˚– 2.84˚) or Rp- (M: 2.11˚, 95% CI: -1.11 – 5.40) conditions.

Experiment 3

We thought that if our experiment showed a revRIF effect on retrievability (as we saw a trend for in Experiment 2), then we would also see the Rp- color reports shift away from Rp+. However, Experiment 2 did not produce such a shift. In considering this, we realized that we had failed to consider some possible outcomes of competition. In our model, a key part of the differentiation process happens when the Rp- item is presented *after* its connections to shared features in the hidden layer have been severed due to competition; according to the model, this is when the Rp- item picks up new features to "replace" the severed shared features, completing the differentiation process (see **Fig 2.1 D, E**). We had failed to consider the possibility that, if the Rp- item's memory is sufficiently weakened by retrieval practice, participants might retrieve the Rp+ item's color when the Rp- item is presented (a swap error), which could result in the representation of the Rp- item being pulled toward the Rp+ item in color space, not away from it (**Fig 2.6;** see also discussion of "asymmetric integration" in Chapter 3). If this happens on a fraction of trials, it would offset (and thus mask) any differentiation that is taking place.

Intuitively, the risk of swap errors is a function of the number of features that are shared between the Rp+ and Rp- item; in this regard, it is important to note that Rp+ and Rp- items have similar shapes—this strong shape similarity could be driving swap errors. Our previous explanation of differentiation in the model (**Fig 2.1**) may have missed this subtlety by not including units representing shape properties. If we additionally consider shape properties,

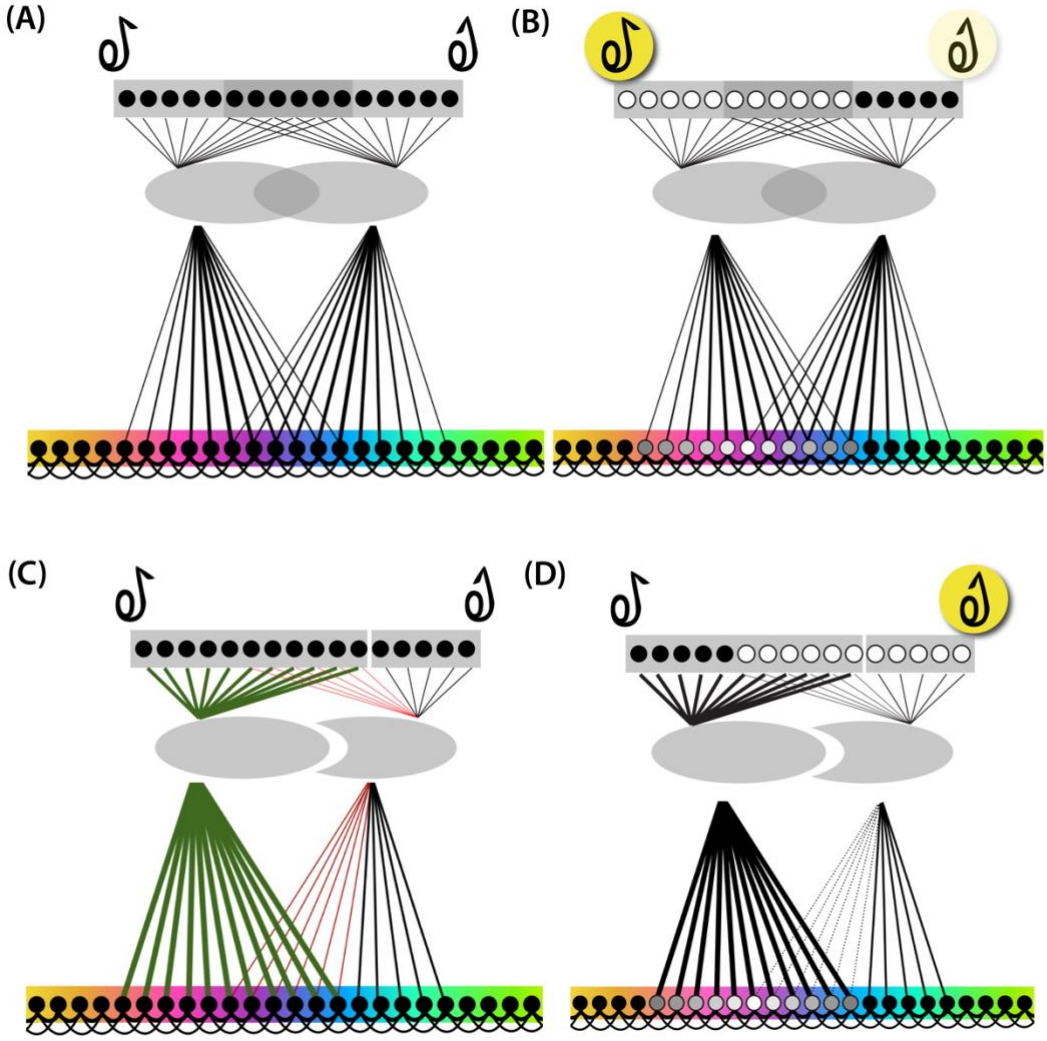


Figure 2.6: Updated theoretical model for revRIF's effect on color, showing how easily swap errors may occur. (A) A continuous input layer is shown in grey at the top, whose features reflect the shape features that are either specific to Rp+, Rp-, or shared. **(B)** During retrieval practice of the Rp+ item, the shape units specific to the Rp+ item and the shared shape units activate (white units). This then activates units in the hidden layer for the Rp+, which in turn activates purple in the color layer. Because of the overlap with the Rp- item, Rp- hidden units activate moderately. **(C)** Because of the high coactivity between the Rp+ units and shared units in all three layers, these connections get strengthened (green lines). The shared units in the shape, hidden, and color layers are all "claimed" by the Rp+ item. Moreover, since retrieval practice caused moderate activation of the Rp- item through spreading activation, the shared units in all three layers sever their connections to the Rp- item (red lines). **(D)** Critically, when the Rp- item is shown again, both the Rp- shape units and the shared shape units activate. These shared units have strong connections to the Rp+ hidden layer units. This causes the pathway for Rp+ to become active, and results in a swap error—the color associated with the Rp+ is recalled, rather than the Rp- item.

highlighting unique shape features of Rp- items (i.e., shape features not shared with Rp+ items) during restudy could prevent swap errors, thereby allowing differentiation to occur. We refer to this idea as "cueing from the side": If you think of a continuum from the Rp+ item's shape (on the left) to the Rp- item's shape (on the right, as in **Fig 2.6**), this cueing strategy involves starting from the far right side of the continuum (i.e., the features that are least similar between the Rp- and Rp+ item) and then moving "inward" toward the Rp+ item.

Methods

25 naïve individuals (21 women, 4 men; mean age: 20 years old) participated. The stimuli and procedure were identical to the previous experiment with the following changes: First, for simplification, we decided to only run Rp+ and Rp- items, without Nrp. Without the Nrp baseline condition, this is no longer a retrieval induced forgetting paradigm, but we can still see if distortion occurs for Rp- by comparing it to Rp+. This also meant we could increase the number of items per condition since all items are either Rp+ or Rp- now (48 items total, 24 Rp+ items and 24 Rp- items). The color difference between pairmates was still 96˚.

The most important change for this study was the implementation of cueing-from-the-side for all trials of the Retrieval Practice / Restudy Task and the Final Recall Task (**Fig 2.7**). We wanted the shape to first appear in such a way as to highlight the individual features, and only gradually show the shared shape-features. Because of the additional time for stimulus presentation, we reduced the number of blocks from eight to six.

Since we included no restudy or correct feedback on the Rp- trials, these trials also undergo a kind of retrieval practice, but we still use the Rp- terminology in order to refer to these trials; we still predicted the Rp- items would distort instead of the Rp+ items, since the correct feedback on Rp+ items provides an anchor to recenter the memory correctly, whereas that feedback was not given on Rp- trials.

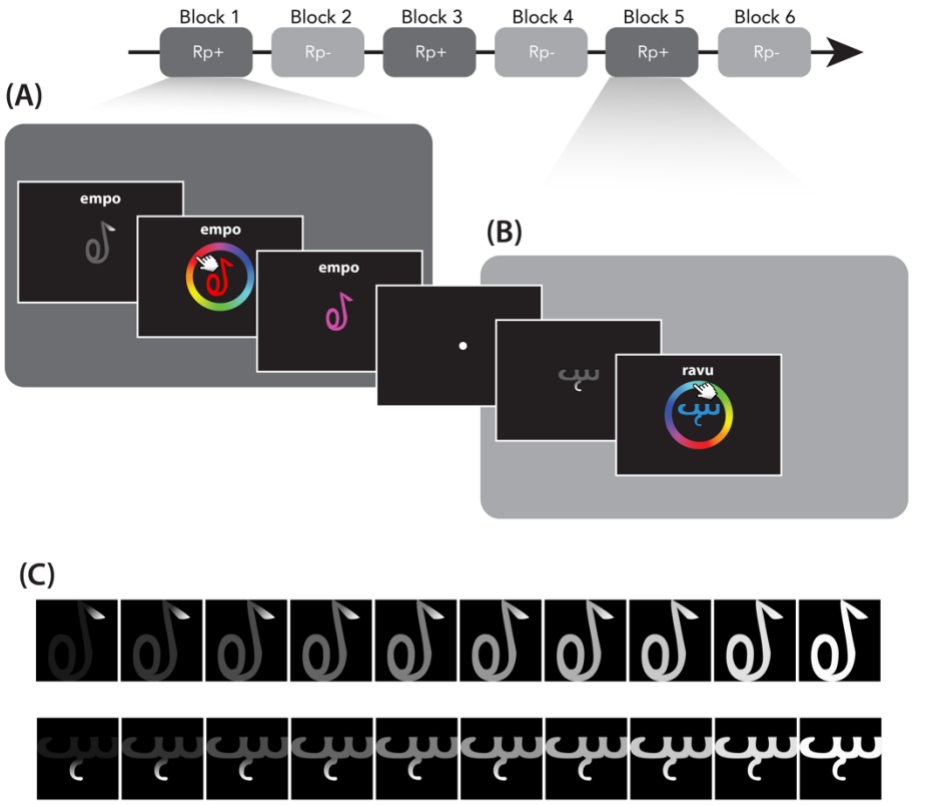


Figure 2.7: Retrieval Practice / Restudy Task for Experiment 3. *The six blocks alternated between Rp+ blocks and Rp- blocks.* ***(A)*** *Procedure for Rp+ Retrieval Practice trials. The shape was presented with the individual features highlighted in white, and the shared features would "fade-in" Then the color wheel appeared and participants made their color response. Once the cursor hovered over a color, the shape was filled with that color. After making their response, the correct color was shown as feedback.* ***(B)*** *Procedure for Rp- Restudy trials. The only difference between Rp+ and Rp- trials was that no correct color was shown in Rp- trials.* ***(C)*** *The fade in was achieved by progressing through ten iterations of shading until all the features were white. The initial level appeared for 1 s and the subsequent nine levels appeared over 1.5 seconds. The final level appeared for an additional 1 s.*

For simplicity, we changed a couple other features of the study. We removed the final Same-Different Task on Day 2. We also were concerned that even the brief 50s restudying of the correct color in the previous experiment helped to reactivate the Rp+ color on Rp- trials, so we removed any color presentation during Rp- trials.

Results and discussion

Results for the model fitting parameters are shown in **Fig 2.8**. There was a difference in *correct recall* probability, in that the Rp+ (M: 66.28%, 95% CI: 58.74% – 73.27%) was remembered better than Rp- items (M: 35.84%, 95%: CI: 16.49% – 52.93%) (p <.05). Since the only difference between Rp+ and Rp- trials was whether the correct color was shown after retrieval practice, it indicates that this feedback was enough to boost memory for the Rp+ items.

The precision estimate, *sd*, was not significantly different between Rp+ (M: 19.74˚, CI: 17.36˚ – 22.68˚) and Rp- trials (M: 22.00˚, 95% CI: 9.91˚ – 41.16˚).

We predicted that cueing-from-the-side would lead the Rp- items to distort away from Rp+ items, since the lack of feedback and highlighting of individual shape features may prevent the Rp+ items from reactivating. However, we found the shift parameter *mu* was not significantly

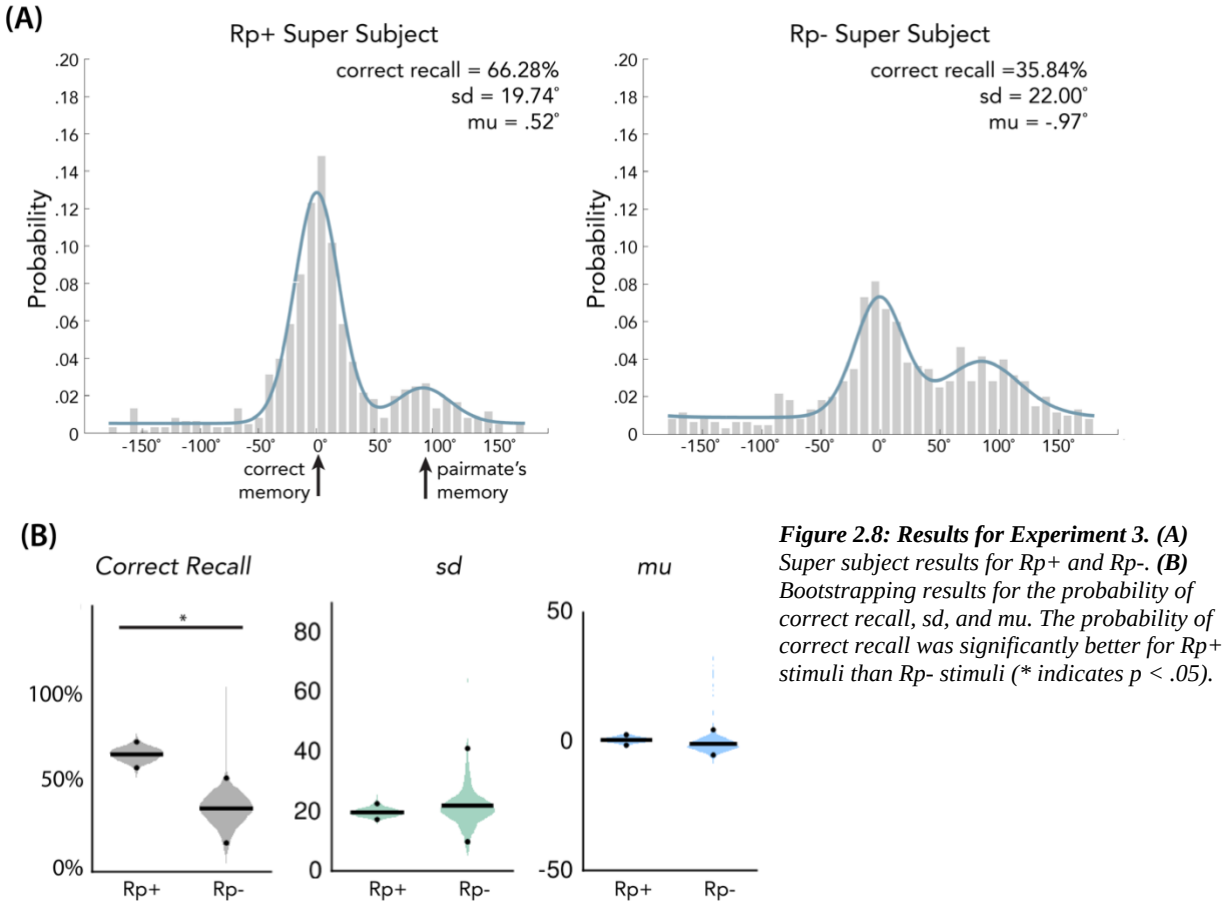


Figure 2.8: Results for Experiment 3. (A) *Super subject results for Rp+ and Rp-.* ***(B)*** *Bootstrapping results for the probability of correct recall, sd, and mu. The probability of correct recall was significantly better for Rp+ stimuli than Rp- stimuli (* indicates p < .05).*

different from zero for either Rp+ (M: .52˚, 95% CI: -1.51˚ – 2.53˚) or Rp- (M: -.97˚, 95% CI: -5.27 – 4.47˚). These were also not significantly different from each other.

We thought cueing-from-the-side would prevent the Rp+ trial from reactivating on Rp- trials, but that is not the case. We can look directly at the *B* parameter, which is the probability of recalling the pairmate's color. The *B* for Rp- (M: 33.37%, 95% CI: 15.91% – 77.16%) was significantly higher than the *B* for Rp+ trials (M: 13.01%, 95% CI: 8.99% – 17.49%) ($p < .05$). If the cueing-from-the-side worked to prevent the Rp+ pathway from activating, then we would not see a difference in the probability of recalling from the pairmate distribution between Rp+ and Rp-. Since participants were recalling the Rp+ color 33% of the time on Rp- trials, it indicates the shared features still activated often.

We still think the cueing-from-the-side method has merit, but this implementation of the procedure was not sufficient to eliminate the swap errors in the Rp- trials, which is a necessary step before any shifts in Rp- items can occur.

Experiment 4

For all experiments from this point on, we switched to collecting data online using Amazon Mechanical Turk (MTurk). Since the shift in *mu* may be very small and memory reports are imprecise, running the study online provides the benefit both of increasing the number of participants per study, and allows us to rapidly test small changes to the procedure.

However, there are downsides to running a study like this on MTurk, most of all the lower accuracy on tasks. Even when we ran versions that were very similar to the ones we ran in-person, people did not provide high fidelity memory reports, and instead were guessing at much higher rates. We ran several batches of participants in slightly varying versions of the experiment

outlined above until we found one where the probability of correct recall on the final color recall task (as estimated using the mixture model) was above 50%.

We added several simplifications to the study to compensate for the worse online performance. Even so, the basic structure is not different from before, so there was no reason to expect that distortion effects would occur this time when they had not previously. Instead, our hope was to run a study where correct recall was high enough to serve as a baseline measure. Then, each subsequent study from this point on adjusts the paradigm very slightly in order to see

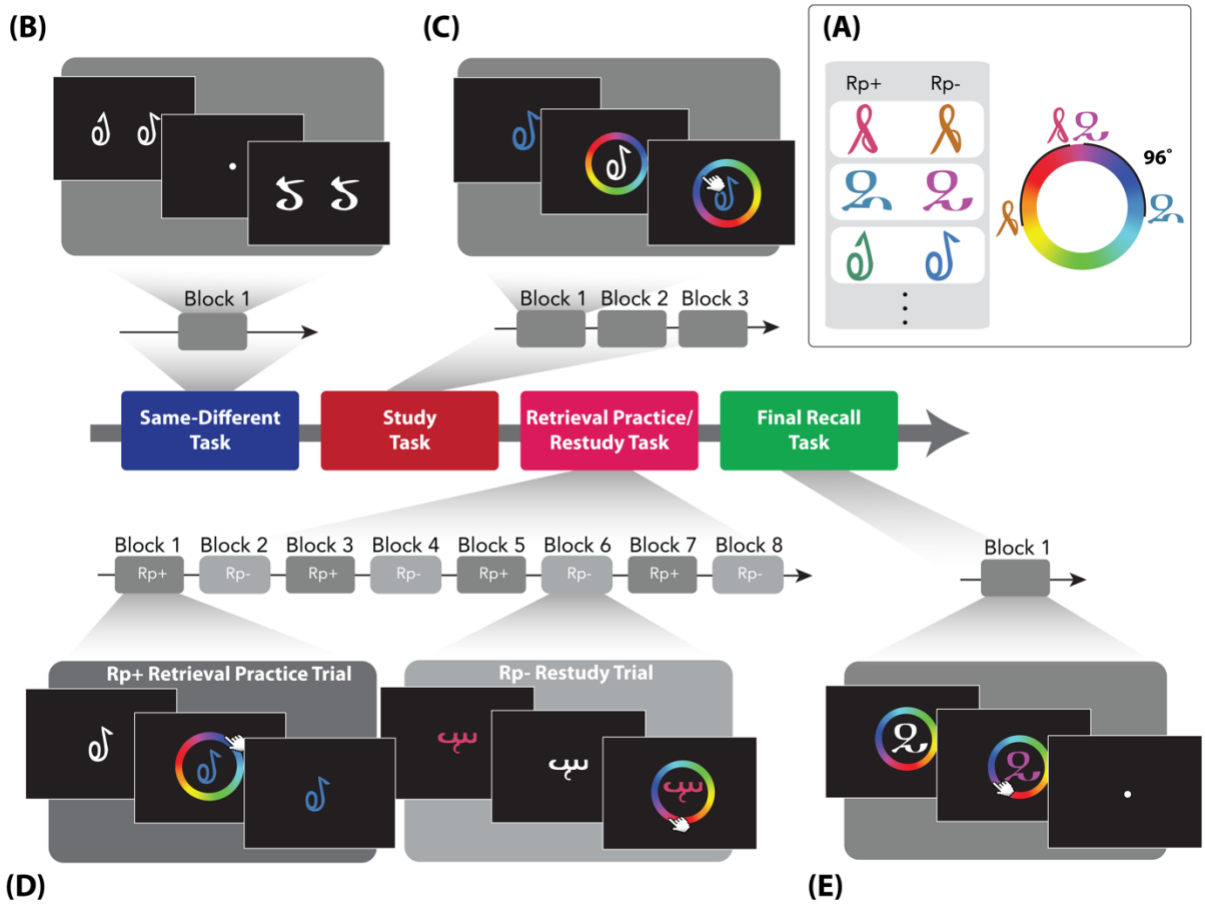


Figure 2.9: Procedure for Experiment 4. ***(A)*** *We reduced the number of stimuli by removing the Nrp condition, and cutting the number of stimuli to 12 items (six pairs). We also dropped all names for the stimuli.* ***(B)*** *The Same-Different Task was the same as before, except we reduced the number of blocks to only one.* ***(C)*** *The Study Task presented all the items once per block, for three blocks total. We changed the task during this phase—after studying the color of a shape, the participants were immediately cued to recreate the color of the shape they just saw.* ***(D)*** *The Retrieval Practice/Restudy Task was unchanged except that there was no Nrp item, so blocks alternated between Rp+ retrieval practice and Rp- restudy. We also doubled the number of presentations per block, so each stimulus was shown twice in a block.* ***(E)*** *The Final Recall Task was unchanged.*

how those changes affect memory reports, and whether we could introduce a change that would induce distortions of pairmates away from each other.

Methods

48 naïve individuals (28 male; 20 female) participated in this experiment. All participants in this study and all subsequent MTurk studies were between 20 and 38 years old. While the task was now run online rather than in person, the procedure was similar to Experiment 1 with several simplifications (**Fig 2.9**). First, we dropped the names from the stimuli, and dropped the Name Learning task. We reduced the number of blocks in several of the tasks, and we reduced the number of stimuli to 12 items (six pairs). We dropped the Nrp condition as we did in Experiment 3, since that cuts down on the memory load and still allows us to investigate shifts in the color memories (however, unlike Experiment 3, the Rp- items still undergo color restudy in this experiment, for 2000 ms). With these changes, we were able to make this a one-day experiment lasting approximately 30 minutes.

To ensure that participants were paying attention to the colors at study, we changed the Study Task. Rather than report how vividly they were able to imagine the color, participants were prompted to record the color of the shape they just studied.

Results and discussion

Results for the model fitting parameters are shown in **Fig 2.10**. The *correct recall* parameter was not significantly different between the Rp+ (M: 49.88%, 95% CI: 34.78% – 64.20%) and Rp- trials (M: 49.77%, 95% CI: 37.24% – 61.18%), but there was a difference in

precision—the *sd* for Rp+ trials (M: 17.9˚, 95% CI: 11.97˚ – 25.22˚) was lower than the *sd* for Rp- trials (M: 26.75˚, 95% CI: 19.90˚ – 36.30˚) (p <.05).

There was no shift away from zero for either Rp+ or Rp- trials, indicating no distortion effects, and none of the shifts were different to each other for any of the conditions: The shift (*mu*) in the Rp+ trials was -2.86 (95% CI: -6.77˚ – 1.00˚), and the shift (mu) in the Rp- trials was -3.30 (95% CI: -9.63˚ – 2.10˚).

We were not expecting a shift in the color memory reports given that there is no major difference here compared with previous studies we ran. While the correct recall percentage may be lower than when we ran this in person, it is high enough for both Rp+ and Rp- that we felt this was a good baseline to explore what may trigger distortion in the subsequent studies.

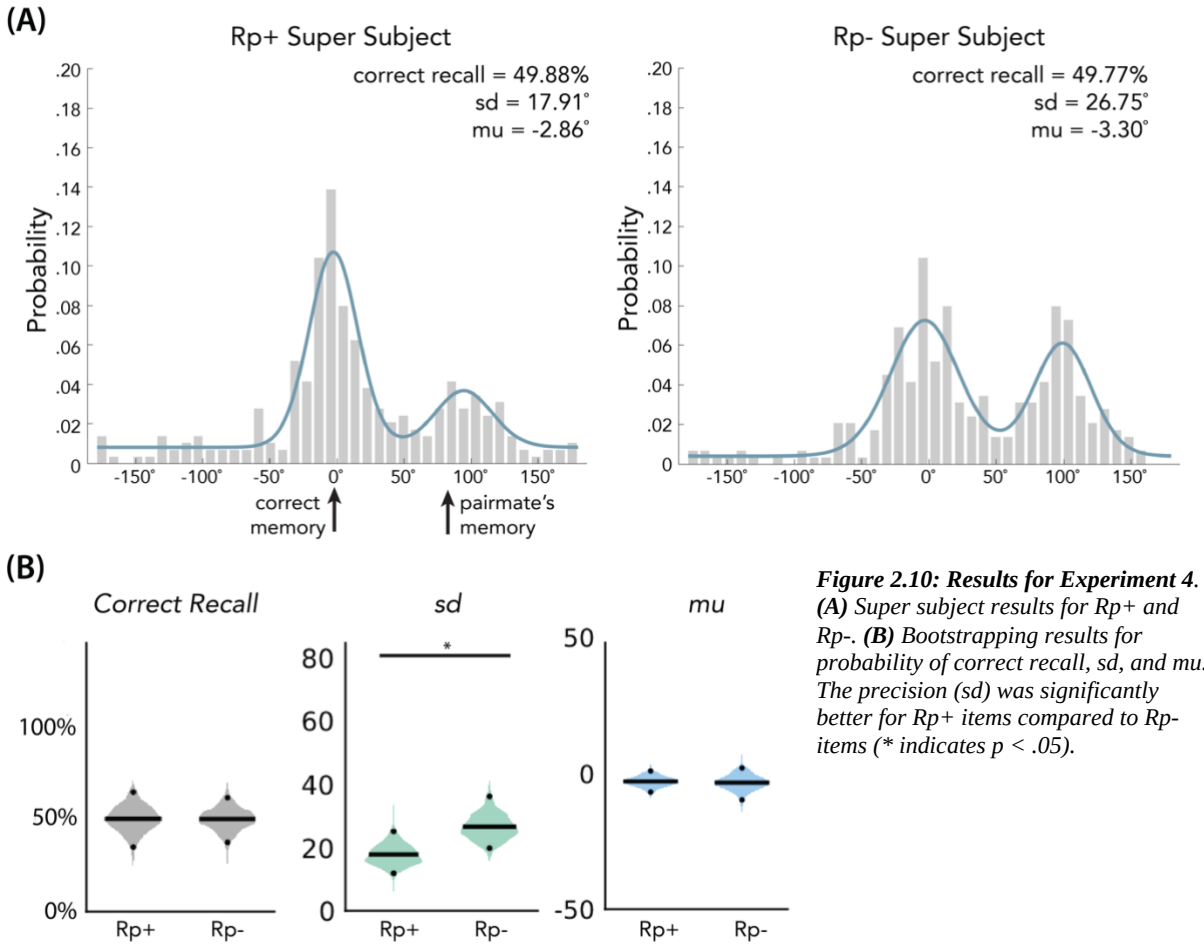


Figure 2.10: Results for Experiment 4. ***(A)*** *Super subject results for Rp+ and Rp-.* ***(B)*** *Bootstrapping results for probability of correct recall, sd, and mu. The precision (sd) was significantly better for Rp+ items compared to Rp- items (* indicates p < .05).*

Experiment 5

The first change we tested was adding a one-day break between the Retrieval Practice/Restudy Task and the Final Recall Task so that participants could sleep. Several studies that show representational change or weakening of competitors happen over a two-day span [10,66], and it is possible that the inclusion of sleep is critical. Neural activity is replayed during sleep leading to improved memory [95–97]. Moreover, competition between memories can modulate how they are consolidated during sleep [98], and some evidence suggests that the Rp- item can be repaired through consolidation during sleep [99]. According to one model of REM sleep [100,101], these reparative effects occur during REM sleep because it provides a period of interleaved learning of competing memories.

It is possible that the paradigm we used was enough to create weakening of the Rp- item, but sleep is needed to restore the connections of the Rp- item and strengthen its connections to the new units that would cause the color report to distort away from Rp+. Of primary interest to us was whether color reports of Rp- items would be distorted away from Rp+ items following sleep, but we were also interested in whether testing color memory on a subsequent day would worsen memory reports or not. If color memory is not weakened significantly, then even if this procedure does not create distortion effects, it may be helpful to keep the one-day break when testing further versions of the experiment (to allow for the possibility of sleep-related plasticity effects). We also compared fitted parameters in this study to the previous, one-day study.

Methods

32 naïve individuals took part in this study (23 men, 9 women). The procedure in this study was identical to the previous study (Experiment 4), except that we added a break (24 hours

+- 4 hours) before the Final Recall test. We also compared each fitted parameter of the mixture model to the same parameter in the previous study by comparing the 95% confidence intervals.

Results and discussion

The results of the mixture model fit to the data are shown in **Fig 2.11**. There was no difference between the *correct recall* probability in the Rp+ trials (M:44.74%, 95% CI: 30.49% – 65.21%) compared to the Rp- trials (M: 32.24%, 95% CI: 17.96% – 47.15%). The *sd* estimate also was not different between the Rp+ trials (M: 21.17˚, 95% CI: 14.90˚ – 42.00˚), and Rp- trials (19.66˚, 95% CI: 9.95˚ – 35.02˚). Additionally, when comparing these parameters to those

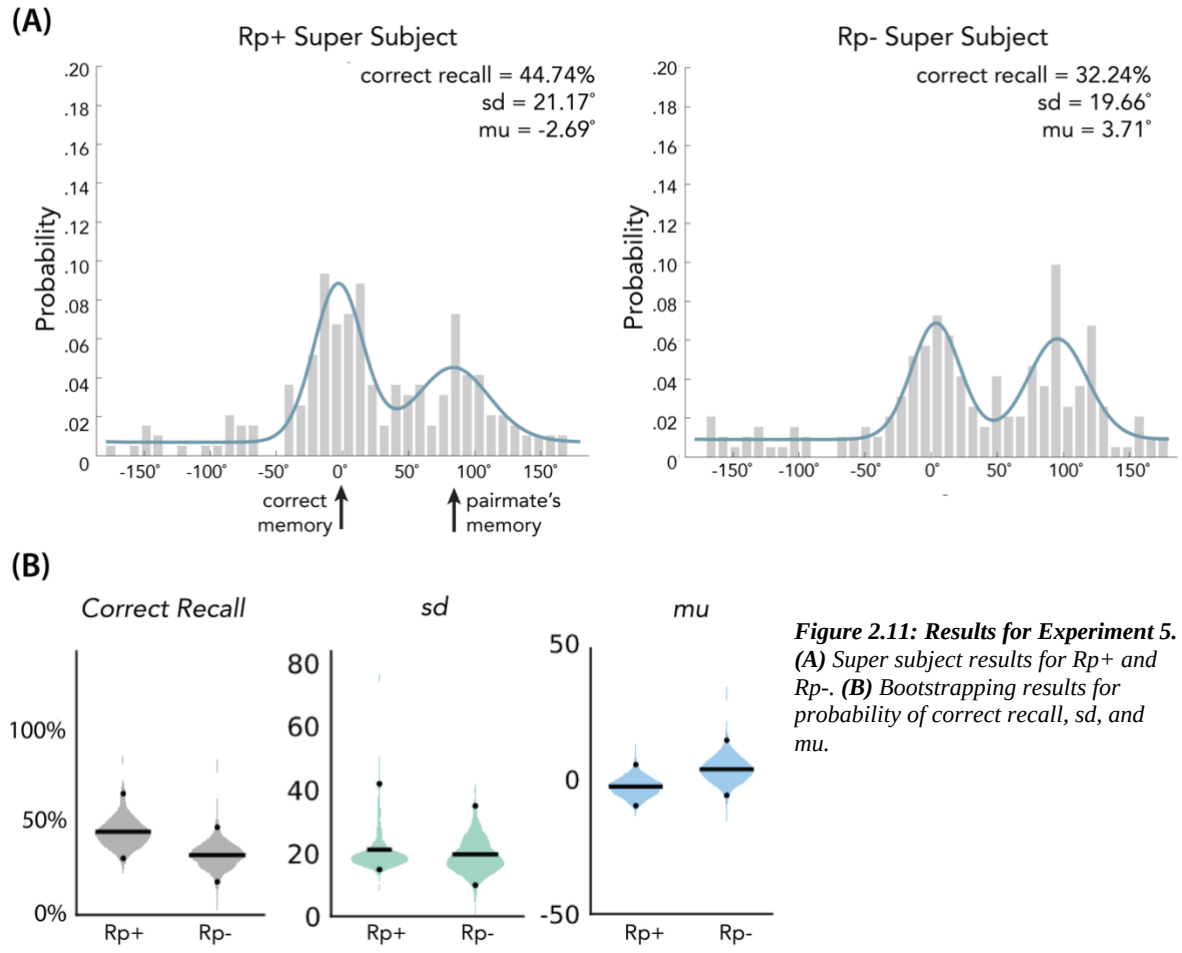


Figure 2.11: Results for Experiment 5. ***(A)*** *Super subject results for Rp+ and Rp-.* ***(B)*** *Bootstrapping results for probability of correct recall, sd, and mu.*

fit in the previous study, none of them were significantly different, indicating that delaying the final color report by a day did not have a significant impact on color memory reports.

We were interested in whether adding a day break might induce a shift of Rp- items away from their pairmates, but this did not occur—the *mu* for Rp+ (M: -2.69˚, 95% CI: -9.74˚ – 5.44˚) and Rp- (M: 3.71˚, 95% CI: -5.93˚ – 14.38˚) were both not significantly different from zero.

While this study did not show distortion, we decided to continue including a one-day break, since there may be benefits of sleep for consolidation for representational change, and our results indicate that this change did not hurt memory performance.

Experiment 6

When we began this series of studies, we were not aware of any study that investigated how competition between similar memories may affect distortions in reports of those stimuli. While we were running the studies above, we learned of a study investigating these effects in color space. In Chanales et al. (2021) [102], participants learned to associate objects and faces. Objects were identical except for their color values. The researchers found that when the color similarity between pairmates was high enough, a "repulsion" effect occurred, where the colors reported were distorted away from the competitor. This study provides support for the theory that motivated our investigation.

Why did Chanales et al. (2021) [102] find this effect in color memories when we did not? There are a number of differences in the paradigms used. For one, in the Chanales et al. (2021) [102] task, participants underwent interleaved study and recall practice. The recall trials could be either a) color recall trials (where the object and associated face were given and the subject had to identify the correct color), or b) face recall trials (where the color and object were given and

participants had to identify the correct face). The latter task can be thought of as a "backwards cueing" task, in that the feature of interest (color) is used to cue the feature used to discriminate between pairmates (face). It is possible such a task is important for making participants attend to the features that are different between the objects—in Chanales et al. (2021) [102], that feature is the individual face, and in our study, that feature is the non-shared shape elements. As we discussed earlier, it may be important to provide extra nudges to make sure participants are paying attention to unique features at retrieval (otherwise swap errors can occur; see **Fig 2.6**). A backwards cueing task may be important for this.

In the context of our study, an analogous backwards cueing task would be one where the color is shown and participants choose the shape. When considering how to implement this task, we wanted to preserve the "retrieval practice" logic of our study, whereby the Rp+ item triggers a competitive retrieval process (resulting in the Rp- item coming to mind and being repelled away from the Rp+ item) but retrieval of the Rp- item is not competitive. As described below, we decided on a compromise—including backwards cueing (to promote attention to unique features) on study trials but making it easy enough so it would not trigger competitive retrieval.

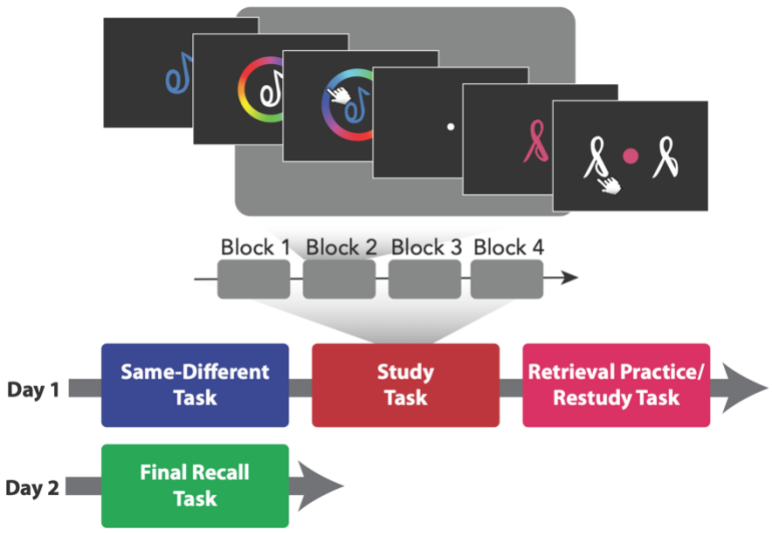


Figure 2.12: Procedure for Experiment 6. *The tasks were identical to the previous experiment except the Study Task. In the previous experiment, all study trials were immediately followed by a prompt to report the color of the shape just shown (i.e., the second and third screens above). Now, study trials could be followed either by these color recall trials, or shape recall trials, where the colored shape is immediately followed by a color cue and participants had to report the associated shape (the fifth and sixth screens above). Both pairmates were shown on the left and right of the central color, and participants used their mouse to click their recollection of the correct item. In this study, each item was still only shown once per block, and color reports versus shape reports for each item were done in alternating blocks. To accommodate this, we increased the number of blocks from three to four.*

Methods

44 naïve individuals (18 male, 26 female) participated in this study. The task was the same as in the previous study (Experiment 5), except for the task during the Color Study phase. Following the presentation of each colored shape in the Study phase, participants were either shown the shape and they had to report the color just shown, or they were shown the color and they had to report the shape just shown (**Fig 2.12**). The short delay between studying an item and being tested on the item's color or shape made this a very easy task—as noted above, our goal was to ensure that participants would attend to the unique color and shape features of each item on study trials, without triggering a competitive retrieval process (we wanted to isolate competition to the Rp+ retrieval practice trials, which come after the study phase). Color tests and shape tests were done on alternating blocks for each item, so that each item was still only shown once per study block. We also increased the number of blocks from three to four so that each stimulus could have a shape test or color test on alternating blocks.

Results and discussion

The results of the mixture model fit to the data are in **Fig 2.13.** There was no significant difference in the proportion of trials estimated to come from the *correct recall* distributions between the Rp+ (M: 49.47%, 95% CI: 37.78% – 61.37%) and Rp- trials (M: 51.33%, 95% CI: 41.22% – 60.39%). The probability of *correct recall* in the Rp- condition was improved from the previous study (increasing from 32.24% to 51.33%) ($p < .05$). Consequently, the changes to the Study Task improved the general retrievability of the items by increasing the likelihood of *correct recall* for Rp- items.

The difference in precision between the Rp+ trials and Rp- trials was significant; *sd* in the Rp+ trials was 13.02˚ (95% CI: 9.86˚ – 16.15˚), but was worse in Rp- trials (M:24.45˚, 95% CI: 19.69˚ – 30.48˚) ($p < .001$). Comparison between this and the previous study indicates that the difference in precision between Rp+ and Rp- is driven by an increase in precision for the Rp+ items ($p < .04$), which in the previous study had a mean precision value of 21.17˚. Consequently, this measure of retrievability was also improved by the changes to the Study Task.

With the increased retrievability compared to the previous study, we were interested in whether there would be any distortion effects in the Rp- condition. However, the model fitting results indicate that there was no distortion (*mu)* in the recalled color for either Rp+ (M: -1.66˚, 95% CI: -4.82˚ – 1.85˚) or Rp- trials (M: -0.71 ˚, 95% CI: -5.33˚ – 7.09˚).

These results provide some evidence that the addition of a backwards cueing task helped participants to discriminate the stimuli. It may be that this provides a more effective approach to highlighting the distinguishing features of the Rp+ and Rp- items than the cueing-from-the-side

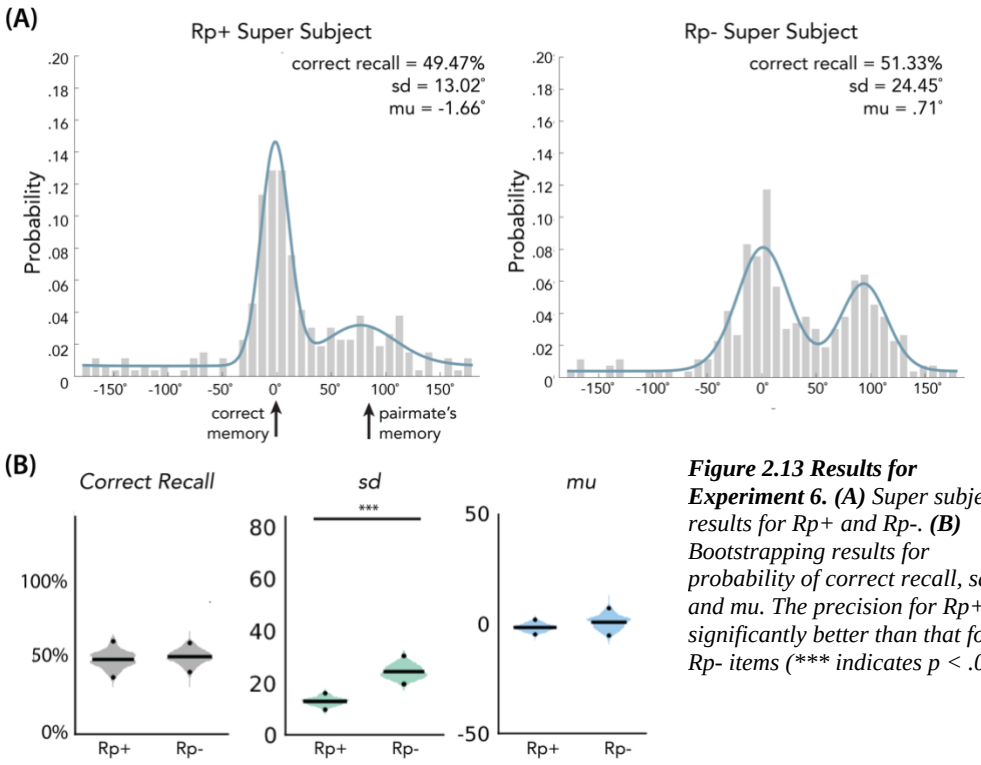


Figure 2.13 Results for Experiment 6. (A) *Super subject results for Rp+ and Rp-.* ***(B)*** *Bootstrapping results for probability of correct recall, sd, and mu. The precision for Rp+ was significantly better than that for Rp- items (*** indicates p < .001).*

approach we tried in Experiment 3. Nonetheless, no distortion effects occurred. A potential reason may be that the pairmates we used did not evoke the appropriate amount of competition to create distortion effects—we explored this possibility in the next study.

Experiment 7

In Chapter 1 we discussed how varying pairmate similarity may affect the outcome of learning. Chanales et al. (2021) [102] provided additional support for this, since they manipulated pairmate similarity. When color difference was low (72˚ and 48˚), memory reports were accurate. When pairmates were more similar (i.e., when color difference was decreased to 24˚), color reports were distorted away from pairmates. When pairmates were even more similar (i.e., when color difference was decreased to 6˚), that distortion effect was reversed (we discuss the mechanism for this adaptation in detail in Chapter 3). The color difference we had been using was 96˚, less similar than any condition in Chanales et al. (2021) [102]. While task demands may affect the similarity needed to produce distortion effects, we wanted to match the condition in Chanales et al. (2021) [102] to see if this was sufficient to trigger distortion of color memories.

Methods

43 naïve individuals (19 male, 23 female, 1 unreported) took part in this study. Other than the reduction in color difference (now 24˚), the procedure was identical to the previous study.

Results and discussion

The mixture model fit to the super subject data is shown in **Fig 2.14**. On the surface, accuracy seems to be much higher in this study compared to the last for both Rp+ (M: 92.09%, 95% CI: 83.53% – 98.49%) and Rp- (M: 91.66%, 95% CI: 84.28% – 96.71%) ($p < .001$). However, making the task more difficult by increasing the similarity between pairmates should not increase retrievability. Furthermore, the probability of drawing from the competitor's distribution (*B)* in the Rp- condition lowered from 33.71% of trials in the previous study to only 1.36% of the trials on this study ($p < .001$). It would be expected that making the items more confusable should *increase* the swap rate, not lower it.

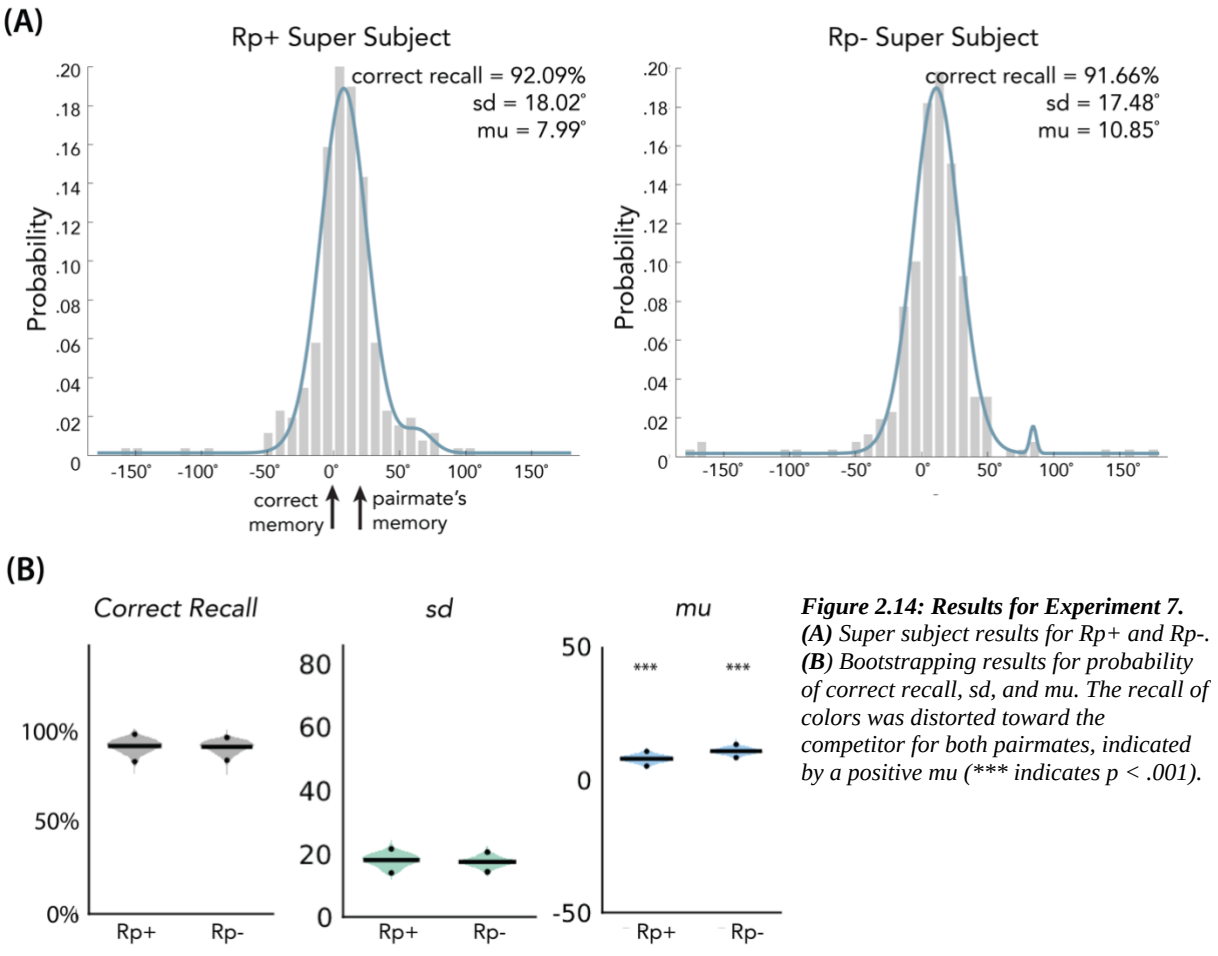


Figure 2.14: Results for Experiment 7. ***(A)*** *Super subject results for Rp+ and Rp-.* ***(B)*** *Bootstrapping results for probability of correct recall, sd, and mu. The recall of colors was distorted toward the competitor for both pairmates, indicated by a positive mu (*** indicates p < .001).*

It may be that the new color difference makes model fitting more difficult, considering the relatively low precision and high color similarity. The *sd* estimate in the Rp+ trials was 18.03˚ (95% CI: 13.99˚ – 21.59˚) and in the Rp- condition was 17.49˚ (95% CI: 14.34˚ – 20.54˚). Since the "correct" distribution and the "bias" distribution are so close and the precision is relatively low, the model may be considering actual swap errors as correct recall responses.

Additionally, since precision is low and color similarity is high, it is possible participants are treating both colors as one, and are motivated to report an in-between color for both pairmates. In effect this means they receive double the presentations compared to the last experiment, if they are only encoding one color per pair. This would explain why the probability of *correct recall* is so much higher than before. This is also supported by the lower overall guess rate (*g*) for all trials in this study (6.45%) compared to the previous study (21.58%; $p < .05$)

We had predicted that increasing the color similarity may put pairmates into enough competition that Rp- items would distort, but that is less likely given the issues described above. Unsurprisingly, then, the model showed that items were distorted *toward* each other rather than away, since the *mu* was positive for both Rp+ (M: 7.99˚, 95% CI: 5.26˚ – 10.76˚, $p <.001$) and Rp- (M: 10.85˚, 95% CI: 8.45˚ – 13.37˚, $p <.001$), again supporting the argument that participants may only be encoding one color per pair.

Experiment 8

In the previous study, we increased pairmate similarity to increase competition between pairmates, but precision may be too low for the model to accurately separate correct recall errors from swap errors. If precision could be increased, our mixture model might be able to fit the data better. Increasing the ability to recall the correct item might also be helpful—the high probability

of correct recall in the previous study may be a result of the merging of the target and competitor bumps, rather than actually recalling the correct item.

To do this, we decided to use more retrieval practice. Previously, retrieval practice only happened in the Rp+ trials. Since retrieval practice can improve learning, we decided to remove the Rp+ / Rp- distinction, so that all trials could get the benefit of retrieval practice.

Methods

42 individuals (17 male, 21 female, 4 unreported) participated in this study. While the stimuli remained unchanged, the procedure of this study had the most significant changes from previous versions (**Fig 2.15**). All stimuli were treated the same, eliminating the Rp+ / Rp- distinction. On the first day, participants completed eight blocks of interleaved study-recall tasks. During the Study Task, each item was shown once per block, and participants were instructed to study the stimuli. During recall rounds, participants had to recall either the object's color or shape. Feedback was shown in the shape trials (by a box surrounding the correct shape), but not the color recall trials, in order to help participants learn to discriminate. We did not include feedback on the color recall trials to help allow any

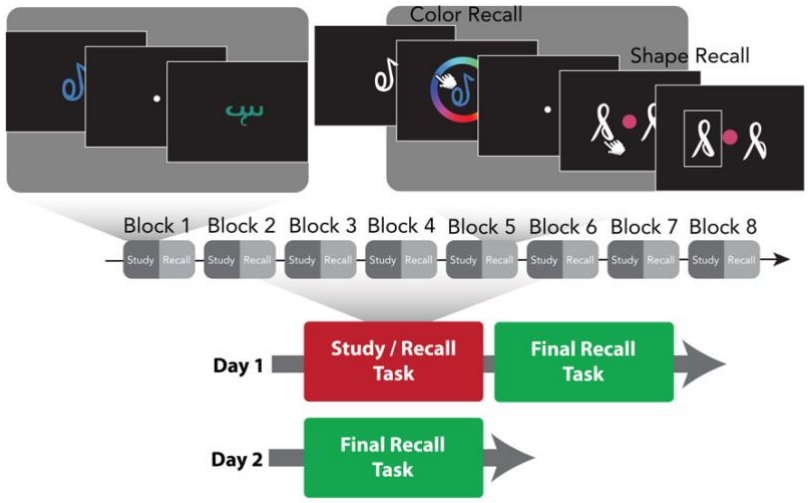


Figure 2.15: Procedure for Experiment 8. *On Day 1, participants completed eight blocks of interleaved study and recall rounds. During the study round, each of the twelve shapes was shown once per block, and participants had no task other than studying the stimuli. During each recall round, each item was shown twice—one time, the participant was cued to recall the color (if the shape was shown), and the other time cued to recall the shape (if the color was shown). We showed correct feedback during the shape recall trials with a box surrounding the correct shape, after the subject's response. We included two Final Recall tests—one at the end of each day.*

distortion effects. We also decided to include the Final Recall test both at the end of the first day and again on the second day.

Results and discussion

Model fitting results for this experiment are shown in **Fig 2.16**. Since this version of the study did not include a distinction between Rp+ and Rp-, we instead looked at the parameters fit for all trials. On the Day 1, the model fitting results show that the *correct recall* probability for all trials was 89.19% (95% CI: 82.05% – 95.27%). The *sd* estimate was 19.21˚ (95% CI: 17.46˚ – 21.48˚); We were interested in whether providing more retrieval practice for all items would increase precision, but this *sd* was not significantly different from that of the previous study.

We were particularly interested in *mu*, which if negative, indicates that items were distorted away from the competitor. Instead, the shift in *mu* on the first day was positive (M: 3.83˚ (95% CI: 1.46˚ – 6.14˚) ($p < .01$). Since the precision was unchanged from the previous task, it may be that participants were still merging the representations of the two pairmates.

On the Day 2, results were similar: correct item recall was 87.02% (95% CI: 80.82% – 93.39%), *sd* was 18.93˚ (95% CI: 17.22˚ – 21.12˚), and mu was positive, at 3.94˚ (95% CI: 1.65˚ – 6.66˚) ($p < .01$). There was also no change in any of these fitted parameters between days.

Consequently, it looks like this study may suffer from the same issues as the previous study, where the high color similarity meant the model may not be fitting the data well—only 4.7% of trials were assigned to the "bias" distribution, so the positive *mu* may simply indicate swap errors or a merged representation. However, this *mu* was significantly less than the positive *mu* in the previous study ($p < .001$), indicating that there may be less merging than before.

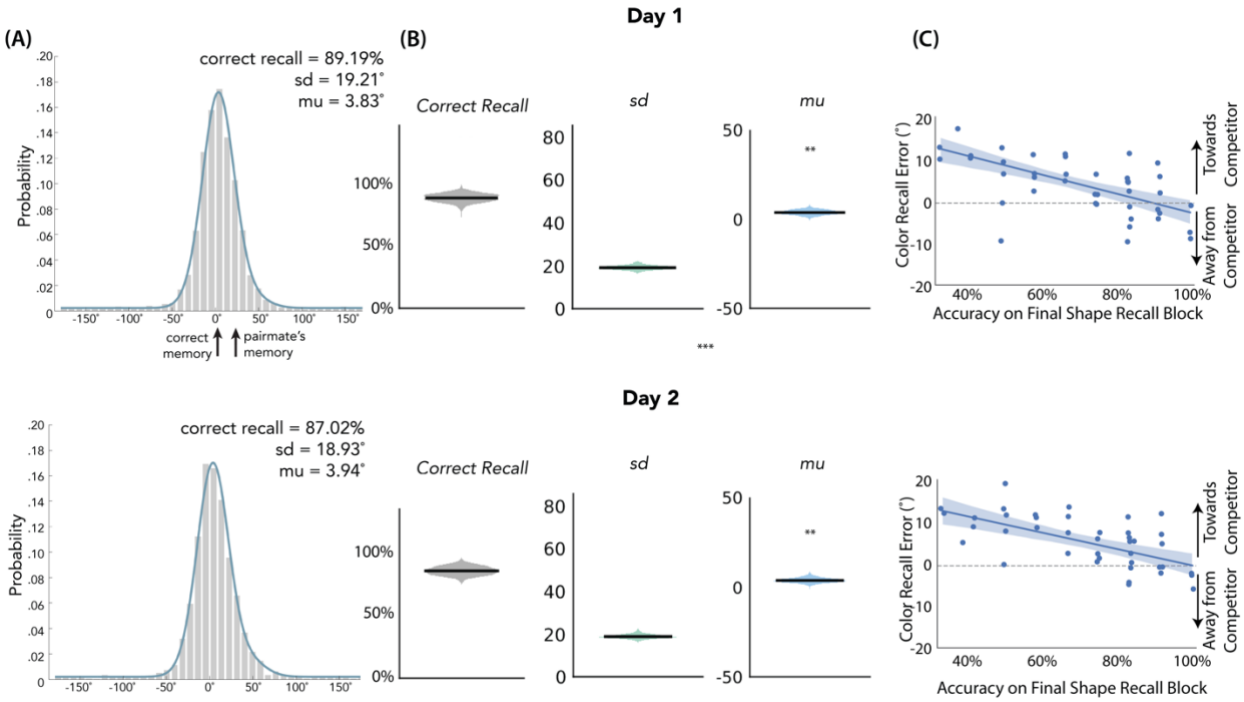


Figure 2.16 Results for Experiment 8: ***(A)*** *Super subject results for all trials for Day 1 (top) and Day 2 (bottom). There was no Rp+ and Rp- distinction in this study, so all trials were pooled.* ***(B)*** *Bootstrapping results for the probability of correct recall, sd, and mu. The recall of colors was distorted toward the competitor for both pairmates, indicated by a positive mu (** indicates p < .01.* ***(C)*** *Color error on the Final Recall Task is shown as a function of the shape accuracy in the final block. The higher the shape accuracy, the lower the average color error (p < .001). Very few participants performed at ceiling for the shape recall task, so it is unclear if the negative slope is a result of these participants recalling with greater accuracy (where color recall error would be at 0˚), or with greater distortion away from the competitor (where color recall error would be negative).*

For this study, we included one more exploratory analysis (**Fig 2.16C**). In general, accuracy on the shape task never reached ceiling; the accuracy of the shape recall trials on the final, eighth block was 71.38%. We thought that accuracy on this task would be important in order to see any distortion effects, but very few of the participants were above 80% accuracy on the shape recall task. We found that the better people were at discriminating the shapes, the less positive their color error was ($p < .001$). However, it is unclear whether that reduction in color error is indicative of people becoming more accurate (where color error would be zero), or if it is more likely to lead to a distortion away from competitors (where color error would be negative).

It is worth noting that this study has many similarities to Chanales et al. (2021) [102]. However, the accuracy on the "backwards cueing" task (where the color was shown and participants had to recall the face) in Chanales et al. (2021) [102], was higher on the last blocks

of learning—it was above 85% for each of the studies in the condition that led to differentiation (24˚). Given that face recall accuracy increased as color shifted more away from the competitor over the course of learning in their study, it is possible the performance on this task is a useful indicator of how well participants are able to discriminate the objects, and consequently, whether repulsion can occur. Perhaps the lower accuracy overall for online participants puts a limit on how well we are able to measure such effects.

General discussion

Lessons

While none of our studies showed conclusive evidence regarding distortion effects, we think there are several important lessons to be drawn from this work.

Selecting appropriate similarity of stimuli

In Chapter 1 we discussed how low levels of similarity may lead to no representational change, moderate amounts of similarity may lead to differentiation, and high amounts of similarity may lead to integration. This suggests that selecting the right level of similarity will be crucial for obtaining color distortion effects.

To maximize the number of trials per condition, we only used one level of color similarity. However, in our first study, we used a color similarity of 64˚. Afterwards, we were concerned this was too difficult for participants, so we decreased the color similarity by raising the difference to 96˚. By decreasing the similarity, we may have made distortion less likely, especially considering the recent studies [102,103] which found this distortion happened with a

color difference of 24˚. In retrospect, we probably should have altered the difficulty in some other way than changing color similarity (e.g., by providing more study trials).

Given the recent studies that found distortion, it may make sense for future studies to start with a color similarity of 24˚. However, surveying the full color spectrum is important; different paradigms and stimuli may affect how items compete, and 24˚ may not be an appropriate difference to consistently lead to distortion. It is possible that the differences in our task made it so 24˚ was still not an appropriate color similarity—in fact, the merging of representations in Experiments 7 and 8 could be a result of the color similarity being too high.

Managing task difficulty

We found our task was very difficult for participants. Ideally, we want participants to recall the correct color most of the time so there are enough trials in the *correct recall* distribution to estimate distortion. Once we switched to online studies, performance deteriorated, and the probability of recalling the correct color hovered around 50% or below.

Some of our adjustments did help improve performance. One of the main improvements came from implementing a "backwards cueing" task, where the color was shown and participants had to select the shape from the two pairmates. Interestingly, this improvement is unlikely to be from extra retrieval practice—we implemented this task in Experiment 6 in such a way that the participants had to report the shape of the stimuli shown immediately prior.

Interestingly, delaying the Final Recall to the following day (Experiment 5) did not lower the probability of *correct recall* significantly. Evidence suggests sleep may be important for repairing damaged competitors [99], so it is useful if this delay does not harm memories.

Moving our study to an online platform had a big impact on performance. We initially reran paradigms much more similar to those we ran in person, but the task was nearly impossible, and very few memory reports were drawn from the correct distribution. We managed to accommodate this somewhat by significantly lowering the demands on memory by decreasing the stimulus number (Experiment 4). But one insight from these studies is that conducting such experiments on competitive dynamics on similar memories may be best in-person.

Limitations on distortion effects

Since we began this line of work, several other studies have found these distortion effects [102,103], and more studies in the future may also. However, distortion effects may be difficult to uncover; Chanales et al. (2021) [102] shows how small adjustments to color similarity can reverse repulsion effects, and in Chapter 1 we discussed how small changes to paradigms have been shown to affect the outcomes of representational change. Identifying boundary conditions on these dynamics can be helpful, and our investigations may be useful to frame future studies.

In the end, Experiment 8 had many similarities to Chanales et al. (2021) [102], but there are still a number of differences between the studies that may explain why we did not find color distortion effects. One of the most obvious differences is the kind of stimuli. We used novel shapes whereas Chanales et al. (2021) [102] used real-world objects. In theory, the predictions of the NMPH should be the same for both, but the novel nature of our stimuli may have presented an added difficulty for participants to overcome.

Furthermore, the method of cueing items was different. We used small differences between the shapes to cue one pairmate or another. Chanales et al. (2021) [102] used identical objects for the two pairmates, and items were cued through an associated face. It could be that

discriminating our stimuli was more difficult. This is supported by the lower performance on our backwards cueing task (where the color was shown and participants had to recall the shape) compared to the backwards cueing task in Chanales et al. (2021) [102] (where the color was presented and participants had to recall the face).

Lastly, several of our studies were run online, which should not impact the predictions of the NMPH in theory, but we did see the ability to recall items deteriorated once we switched. If accuracy is lower, there are fewer trials in the *correct recall* distribution to fit the *mu* parameter.

Implications for future studies

One target for future studies may be to investigate the remaining differences between our study and Chanales et al. (2021) [102]. There are several other potential avenues to investigate.

RevRIF studies

We believe it would still be useful to investigate these effects in a revRIF paradigm, since the NMPH predicts that it will be the Rp- item that distorts (not the Rp+ item; see Chapter 3 for discussion of these asymmetric differentiation effects). We think some of the studies we ran may be useful as a basis for future studies. While the effects were not significant, the results in Experiment 2 indicated the paradigm may produce a revRIF effect on retrievability. However, the color difference in that study was 96˚, and Chanales et al. (2021) [102] indicates these colors may not have been similar enough to get a repulsion effect.

We also think the idea of cueing-from-the-side in Experiment 3 was promising. Swap errors can present a difficulty since they limit the number of trials drawn from the correct

distribution. Additionally, we will discuss in Chapter 3 how reactivating the wrong item on a trial may undo any distortion effects.

There may be other methods to cue-from-the-side more efficiently, for instance, by using stimuli where the individual features can be manipulated with greater control (i.e., artificial shapes that vary along multiple dimensions [104]). The "fade in" could be done by parametrically adjusting the shapes on these continuous dimensions. Another way would be to develop stimuli using pre-trained convolution neural networks that mimic the layers of the human visual system to produce pairs of images that vary in their visual similarity in constrained ways [105]. To cue from the side, you could exaggerate the diagnostic features of those stimuli.

Effects of color similarity on other features

Another goal is to improve our behavioral measure to see if differentiation affects the shape representation, and not just the color representation. One way to measure this is in terms of confusability—if the Rp shapes have differentiated, they should be less confusable than the undifferentiated Nrp shapes. Another approach would be to use a rapid serial visual presentation task developed by Schapiro et al. (2012) [68], which has previously been shown to be sensitive to representational change effects; for instance, a continuous stream of two paired Rp shapes and two paired Nrp could be shown (each item shown an equal number of times), followed by a prompt to respond whether one of the Rp shapes or one of the Nrp shapes appeared more often in the stream. If the Rp shapes have differentiated, they should be less confusable than the Nrp shapes, and people might overestimate the number of presentations of the Nrp item. Another approach would be to look at repulsion of shape characteristics, using stimuli where shape features are varied continuously [106].

Summary

We ran a series of studies to investigate how learning about similar stimuli affects the features of those memories. It follows from the NMPH that such stimuli could lead memories to be distorted away from each other in feature space, if the competing memory activates to a moderate degree (Chapter 1). However, we encountered several difficulties in running these studies. Baseline accuracy was low, and swap rates were high. We tried running several variations on our learning paradigm to help participants focus on the identifying characteristics of each pairmate, but we did not find any conclusive results regarding distortion.

We first ran a series of studies in person, using variations on the revRIF paradigm. One study showed a trend toward the standard revRIF effect on retrievability. This indicates that the stimuli and paradigm we used may be sensitive to revRIF effects, even if we did not adequately put the items into competition to generate a shift in the recalled colors.

We switched to running experiments online around the time we learned of a similar study that did show such color distortion for highly similar memories [102]. We then adapted our study to investigate which aspects of the successful study were missing from ours. We believe one major difficulty we had was the decrease in performance after switching to online studies, but we found various adjustments helped to improve accuracy (for instance, including a "backwards cueing" task to help discrimination, and increasing the amount of retrieval practice trials). While we never saw the distortion effects predicted by the NMPH, we did manage to improve memory performance over the course of these online studies, and our studies suggest various limiting factors on finding such distortion effects. We think the lessons from these studies could provide useful guidance for future studies.

Chapter 3: A neural network model of differentiation and integration

In this chapter, we present a neural network model to provide computational evidence for the theory that the NMPH predicts the representational change discussed in Chapter 1. Intuitions about how a complex system should behave are not always accurate in practice, and verbally-stated theories can contain ambiguities or internal contradictions that are only exposed when one tries to build a working model. Building a model also makes it possible to generate more detailed predictions (i.e., the model can be used to see what follows from a core set of principles). Relatedly, there are likely *boundary conditions* on how a model behaves (i.e., it will show a pattern of results in some conditions but not others). Models can be used to characterize these boundary conditions, which (in turn) can be translated into new, testable predictions.

Consequently, we built versions of this simulation to model three experiments: 1) Chanales et al. (2021) [102], which looked at how the amount of stimulus similarity affected the distortion of color memories; 2) Favila et al. (2016) [10], which looked at the representational change of memories with the same or different predictive consequences; and 3) Schlichting et al. (2015) [11], which looked at how learning curriculum (whether pairs were presented in a blocked or interleaved fashion) affected the representational change of memories in different brain regions. We chose these three experiments because each show surprising results where similar stimuli differentiate or are remembered as less similar than they actually are, and they allow us to explore three different ways in which the amount of competitor activity is modulated.

The model can account for the results of these studies, and it provides several novel insights. Most importantly, the model shows that differentiation must happen quickly (or it will

not happen at all), and that representational change effects are often asymmetric (i.e., one item's representation changes but the other stays the same).

The following is a description of the general model. Later sections describe how we implemented the model separately for the three studies.

Basic network properties

In order to investigate the dynamics of representational change according to the NMPH, we aimed to build a model that could account for both integration and differentiation. Two features of the model were of highest importance to us: 1) The model should only use an unsupervised, U-shaped learning function, and 2) the spread of activity to the competitor should be systematically manipulated. Both these features were necessary since we predicted that the amount the competitor is allowed to activate (and the associated weakening or strengthening after applying the U-shaped learning function to that activation) would determine whether differentiation or integration results. With these features in place, we wanted to be able to apply this model to different paradigms and experimental conditions.

Since we were specifically interested in the way competition affects learning, we decided to focus on the key moment *after* the memories have been formed, when they first come into competition with each other. Consequently, we pre-wired the initial connections into the network for each stimulus, rather than the allowing the connections to self-organize through learning. Doing so meant we could have control over the exact level of competition between pairmates. Moreover, in the interest of keeping the simulations simple, we modeled a single task from each study rather than modeling all of them comprehensively.

While our simulation includes several important simplifications from the studies it is based on, our intention was to create a basic "sandbox" in order to explore the crucial components necessary to lead to differentiation or integration. This is not meant to be the most biologically plausible model—for instance, we do not think that an unsupervised, U-shaped learning function is a replacement for error-driven learning (instead, we think it is a supplementary tool the brain uses to reduce interference of competitors [107]). Nonetheless, we intentionally used no supervised learning in order to explore whether the U-shaped function alone would be sufficient to lead to either differentiation or integration, depending on the competitor dynamics. In the "Discussion" section, we elaborate on the ways this model could be extended in the future.

Model architecture

To investigate how the NMPH could account for both differentiation and integration, we built a neural network model using the Emergent simulation software [108]. All versions of the model have the same basic architecture (**Fig 3.1**; See the following sections for how the model was adapted for each version, and the "Methods" section for a details of all parameters).

The model has four layers: category, item, hidden, and output. The category and item layers are input layers that represent the sensory or semantic

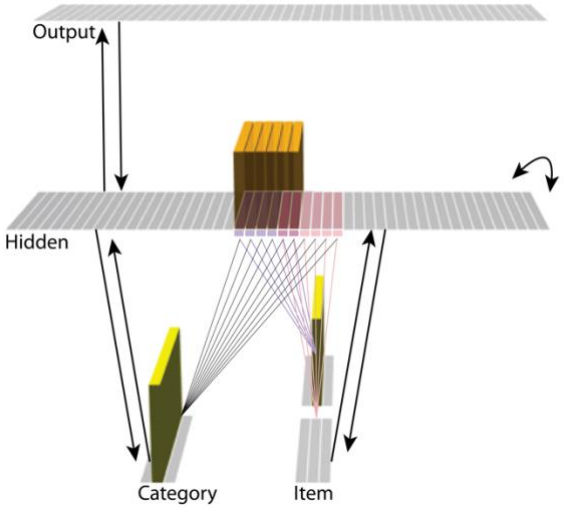


Figure 3.1: Basic network architecture. *The characteristics of the model common to all versions we tested are shown above, with hidden and input layer activity for pairmate A. Black arrows indicate projections common to all versions of the model. All projections were fully connected. Pre-wired, stronger connections between the input A unit and hidden A units (purple) are shown, and pre-wired, stronger connections between the input B and hidden B units (pink) are shown. The category unit is pre-wired to connect strongly to all hidden A and B units. Hidden A units have strong connections to other hidden A units (not shown); the same is true for hidden B units. Pre-wired, stronger connections also exist between hidden and output layers (not shown). The arrangement of these hidden-to-output connections vary for each version of the model.*

features of the presented stimuli. The hidden layer contains the model's internal representation of these stimuli, and the output layer represents either additional sensory features of the input stimulus, or else other stimuli that are associated with the input stimulus. In all models, the category and item layers are bidirectionally connected to the hidden layer; the hidden layer is bidirectionally connected to the output layer; and the hidden layer also has recurrent connections back to itself. All of these connections are modifiable through learning. The model is meant to be as generic as possible, so none of the layers are meant to correspond to a particular brain region (we discuss possible anatomical correspondences in the "General Discussion" section).

Each version of the model learns two pairmates, which are stimuli (binary vectors) represented by units in the input layers. The presentation of any stimulus involves a single unit in both input layers (the category and item layers). Pairmates share the same unit in the category layer, but differ in their item layer unit.

We labeled the item layer units such that the unit in the top row was associated with pairmate A and the unit in the bottom row was associated with pairmate B. Since either pairmate could be shown first, we refer to them as pairmate 1 or 2 when the order is relevant—pairmate 1 is whichever pairmate is shown on the first trial and pairmate 2 is the item shown second.

All projections as described above have weak, randomly-sampled weight values, but we additionally prebuilt some structured knowledge into the network. The input units have pre-wired, stronger connections to a selection of six pre-assigned "pairmate A" hidden units, and six pre-assigned "pairmate B" units. The A and B hidden representations overlap to some degree (for instance, with two units shared, as shown in **Fig 3.1**). We decided to make this hidden layer one dimensional so that the amount of overlap could be easily visualized and interpreted. Laying

it out this way makes identification of pairmate-specific and shared units easier, since the shared units always fall in the middle of the A and B hidden units.

Hidden units representing each individual item are strongly interconnected (that is, the six pairmate A units are highly connected, and the six pairmate B units are highly connected). We also pre-wired some stronger connections between these hidden A and B units and the output units, but the set-up of the hidden-to-output pre-wired connections depends modeled experiment.

Inhibitory dynamics

Within a layer, inhibitory competition between units was enforced through an adapted version of the *k-winners take all* algorithm [39] (see "Methods") which limits the amount of activity in a layer to *k* units. The kWTA algorithm provides a useful way of capturing the "set-point" quality of inhibitory neurons without requiring the inclusion of these neurons directly.

The main method we use to allow activity to spread to the competitor is through inhibitory oscillations; prior work has argued that inhibitory oscillations could play a key role in this kind of learning [1,59,101]. Temporarily lowering the level of inhibition allows competing memories to become moderately active (after which their connections to the target memory are weakened) or even strongly active (after which their connections to the target memory are strengthened). Thus, a larger amplitude to oscillations could lead to more representational change. This theory is consistent with data showing that theta oscillation amplitude at encoding predicts subsequent retrieval success [82–84]. Our model implements oscillations through a sinusoidal function which lowers and raises inhibition over the course of the trial, allowing competitors to "pop up" when inhibition is lower.

Learning

For each projection where learning occurs (bidirectionally between the input and hidden layers, the hidden layer to itself, and the hidden to output layer), we specified the parameters for a U-shaped learning function to allow weakening or strengthening. Once the U-shaped learning function for each projection in each version of the model was specified, we did not change it for any of the various conditions. Details of this function can be found in the "Methods" section.

Competition

Constructing the network in this way allows us to precisely control the amount of competition between pairmates. There are several ways beside amplitude of oscillations to alter the amount of excitation spreading to the competitor. For instance, competitor activity could be modulated by altering the pre-wired weights to force the hidden layer representations for A and B to share more units. Each version of the model (for each experiment) relies on a different method to modulate the amount of competitor activation.

Model of Chanales et al. (2021): Repulsion and attraction of color memories

Key experimental findings

The first experiment we modeled was Chanales et al. (2021) [102]. The study was inspired by the recent neuroimaging studies showing that, after learning, neural representations of similar events can sometimes end up being less correlated than neural representations of unrelated events [10,11,26,31,12,57]. They tested whether a similar "repulsion" effect is observed with respect to how the specific features of competing events are retrieved. In their

experiments, participants learned associations between objects and faces (**Fig 3.2A**)—specifically, participants studied pairs of objects that were identical except for their color value; each of these object "pairmates" was associated with a unique face. Participants' memory was tested in several ways; in one of these tests—the color recall task—the face was shown as a cue alongside the colorless object, and participants were instructed to report the color of the object using a color wheel.

Chanales et al. (2021) [102] found that, for low levels of pairmate color similarity (i.e., 72˚ and 48˚ color difference), participants were able to recall the colors accurately when cued with a face and an object. However, when color similarity was increased (i.e., when the color difference was only 24˚), the color reports were biased away from the pairmate, leading to a "repulsion" effect. For instance, if the pairmates consisted of a red jacket and a red-but-slightly-orange jacket, repulsion would mean that the slightly-orange jacket would be recalled as less red and more yellow than it actually was. When the color similarity was increased even further (i.e., when the color difference was only 6˚), the repulsion effects were eliminated (**Fig 3.2B**).

Potential NMPH explanation

Typical error-driven learning would not predict this outcome, since it would predict that weights should be adjusted to align the guess more closely with the true outcome, leading to no repulsion effects. The NMPH potentially explains the results of this study well, and as the authors note, "the relationship between similarity and the repulsion effect followed an inverted-U-shape function, suggesting a sweet spot at which repulsion occurs."

The amount of color similarity provides a way of modulating the amount of excitation that can flow to the competitor's neural representation. When color similarity is lower (i.e., 72˚

and 48˚), the competing pairmate is less likely to come to mind. Consequently, the competitor's activity will be on the low, left side of the U-shaped function (**Fig 1.4C**). No weakening or strengthening will occur for the competitor, and both pairmate representations remain intact.

However, when color similarity is higher (i.e., 24˚), the additional color overlap means the memories compete more. When the red jacket is shown, the competing red-but-slightly-orange jacket may come to mind moderately, so it falls in the dip of the U-shaped function. In this case, the NMPH predicts that the two items will show differentiation, resulting in repulsion in color space.

Note, however, that when color overlap was raised even further to the point where the pairmates only had 6˚ of difference, the repulsion effect was eliminated. The NMPH could explain this result as the competitor getting so much excitation from the high similarity that it falls on the right side of the U-shaped function, leading to integration (or "attraction" for the behavioral reports of color). But at 6˚ of color separation, there is a limit on how much more similar the color reports could become, and it may not be possible to actually see attraction.

Model set up

Model architecture

To model this task (**Fig 3.2C**), the two input layers (i.e., category and item) represent the colorless object and face associates, respectively. The individual faces were used to cue one pairmate versus another, and so the faces served to identify the individual pairmates. Thus, one stimulus is represented by a unit in both input layers, where two pairmates share the same unit in the object layer (i.e., "jacket"), but differ in the unit for the face layer (i.e., "face-for-jacket-A"

and “face-for-jacket B”). The output layer represents the color-selective units. Units that are closer to each other can be thought of as representing colors that are more similar to each other.

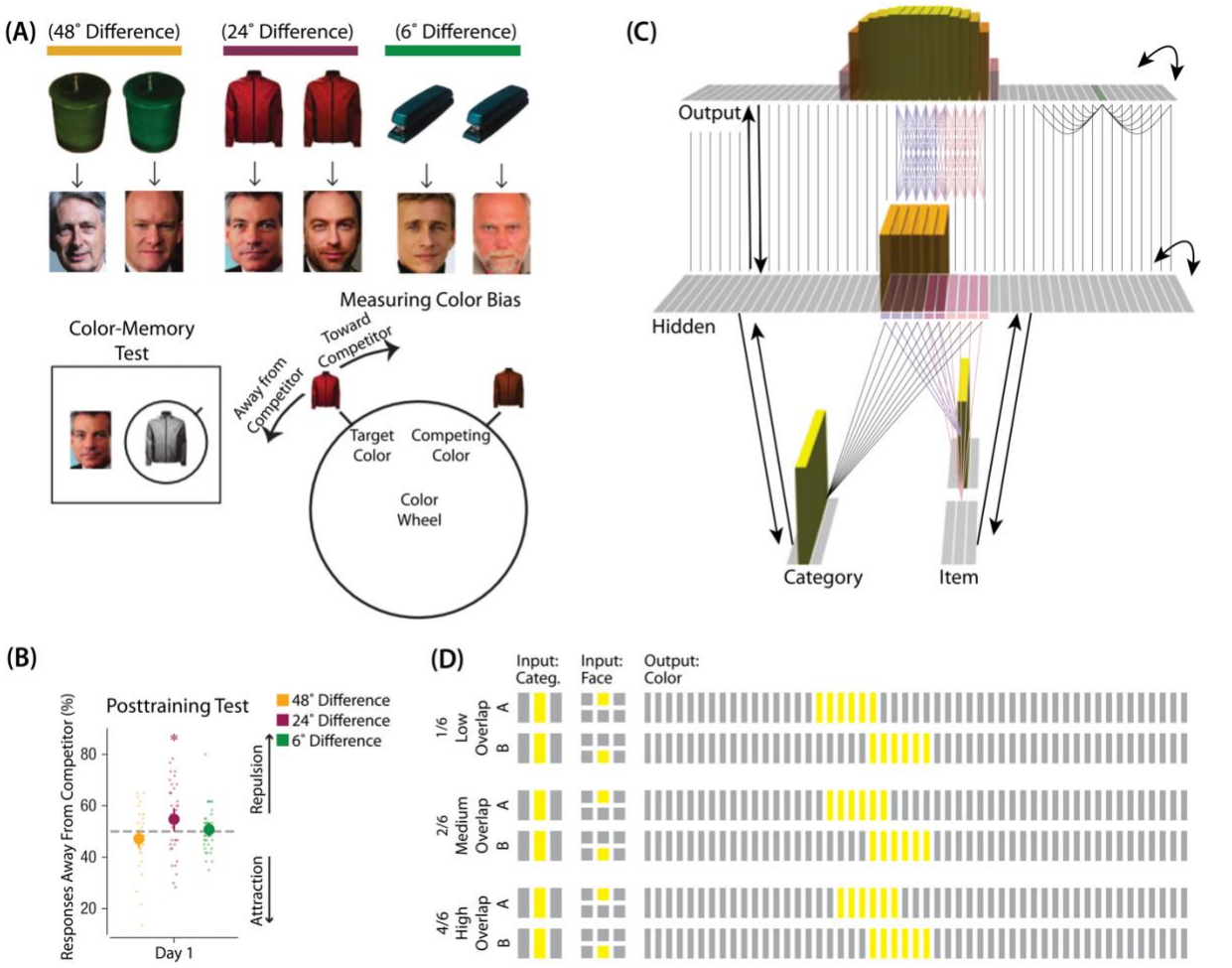


Figure 3.2: Modeling Chanales et al. (2021). (A) *Participants in Chanales et al. (2021) learned to associate objects and faces. The objects consisted of pairs that were identical except for their color value, and the difference between pairmate color values was systematically manipulated to adjust competition. Color memory was tested in a task where the face and colorless object was given as a cue, and participants had to use a color wheel to report the color of the object. Color reports could be toward the competitor (+) or away from the competitor (-).* ***(B)*** *When color similarity was low (48˚), color reports were accurate. When color similarity was raised to a moderate level (24˚), repulsion occurred, and color reports were biased systematically away from the competitor. When color similarity was raised further (6˚), the repulsion effect was eliminated.* ***(C)*** *To model this study, we used the network structure described in “Basic Network Properties,” with the following modifications: this model additionally has a non-modifiable recurrent projection in the output layer, to represent the continuous nature of the color space: each output unit was maximally pre-wired to connect to the seven units on either side of it (one such set of connections is shown to the output unit colored in green), and background output-to-output connections (outside of these seven neighboring units) were set up to be low level and random. The hidden layer additionally has one-to-one pre-wired connections with units in the output layer in the same topographical arrangement. We also maximally connected each of the six hidden A units to six pre-assigned A units in the output layer (purple lines). The same arrangement was made for the hidden B units (pink lines). Activity is shown after pairmate A is presented—the recurrent output-to-output connections let activity spread to units on either side.* ***(D)*** *We included six conditions in this model, corresponding to different numbers of shared units in the hidden and output layers. Three conditions are shown here. The conditions are labeled by the number of hidden/output units shared by A and B. Thus, one unit is shared by A and B in 1/6, two units are shared by A and B in 2/6, and so on. More units overlapping represents pairmates with higher color similarity. We included overlap types from 0/6 to 5/6.*

Knowledge built into the network

As described in "Basic Network Properties," each of the two input face units is pre-wired to connect strongly to the six corresponding hidden units (either the six hidden A units, or the six hidden B units). In this version of our model, we added several extra pre-wired connections, specifically, recurrent output-to-output connections (although we did not include learning for this projection). Neighboring units were pre-wired to have stronger connections, to instantiate the idea that color is represented in a continuous way in the brain [91,90].

The hidden A and B units were maximally connected to six A and six B units in the output layer, respectively. Consequently, the amount of overlap was exactly matched for both the hidden and output layers. Additionally, each non-pairmate unit in the hidden layer had one-to-one connections with a single unit in the output layer. We arranged these one-to-one connections to line up so the left-most hidden layer unit was connected to the left-most output layer unit, and so on. Arranging the units in this way means it is easier to interpret the types of distortion that results in the hidden layer.

Manipulation of competitor activity

The main difference between model versions is the way in which competitor activity is modulated. In this experiment, competition is manipulated through the level of color similarity in each condition. We instantiated this by adjusting the level of overlap between the hidden (and output) color A and B units. One advantage of modeling is that, since there is no constraint on experiment length, we were able to sample a wider range of overlap types than the actual study, which was limited to three conditions per experiment. Instead, we used six overlap types in the model. For all conditions, the two pairmates were each made up of six assigned units in the

hidden layer. However, we could pre-wire the weights from the input layers to the hidden layers so that the two pairmates differed in the number of hidden layer units that were shared (and since the overlap type was also reflected in the output layer, the hidden-to-output connections were pre-wired accordingly as well).

We labeled the conditions based on the number of units that overlap and the number of total units per pairmate (which was always six). A 0/6 overlap means that the two pairmates are side-by-side but shared zero units. A 1/6 condition means they share one unit out of six each, a 2/6 overlap type means they share two units out of six each, and so on. Consequently, the 0/6, 1/6, and 2/6 conditions are analogous to the 72˚, 48˚, and 24˚ color difference conditions in the original study, respectively. The highest similarity type we tested (5/6) is analogous to the 6˚ color difference condition.

Task

The task simulated in the model is a simplified version of the paradigm used in Chanales et al. (2021) [102], since our main interest was what happens at the moment the pairmates first compete, rather than simulating each study fully. Consequently, the training task reflects the color recall task, where the colorless object was shown alongside the face cue, and the participant had to report the object's color on a color wheel. To model this, the external input is clamped to the network (object and face units), and activity is allowed to flow through the hidden layer to the color layer so the network can make a color "guess." We ran a test epoch after each training epoch so we could track the representations of pairmate A and B over time.

As described in "Basic Network Properties," inhibitory oscillations allow the next-most-excited units to pop up toward the end of the trial. This is important to allow potential competitor

units to activate. There is no "correct" color shown to the network, and all learning is based purely on an unsupervised U-shaped learning function of the coactivity of presynaptic and postsynaptic units (see "Methods" for parameter details).

Results

Repulsion and attraction

Chanales et al. (2021) [102] showed that, as color similarity was raised, a repulsion effect occurred where the colors of pairmates were remembered as less similar to each other than they were in reality. When color similarity was raised even further, this repulsion effect went away. In our model, we expected to find a similar U-shaped pattern as overlap was increased.

Our model showed this pattern as stimulus similarity was raised: When the overlap is low, the color memories are stable over learning (see **Video 1**), since very little competitor pop up occurs that could lead to representational change. When color similarity is increased, the competitor pop up is enough for competitor-to-shared connections to fall into the dip of the U-shaped function and sever, leading to repulsion (see **Video 2**). When the similarity is increased further, pop up is high enough for the competitor-to-shared connections to fall to the right of the U-shaped function and strengthen, leading to attraction (see **Video 3**).

To measure repulsion, we operationalized color memory report as the center of mass of activity in the color layer at test. Since for this version of the model the topography of the output

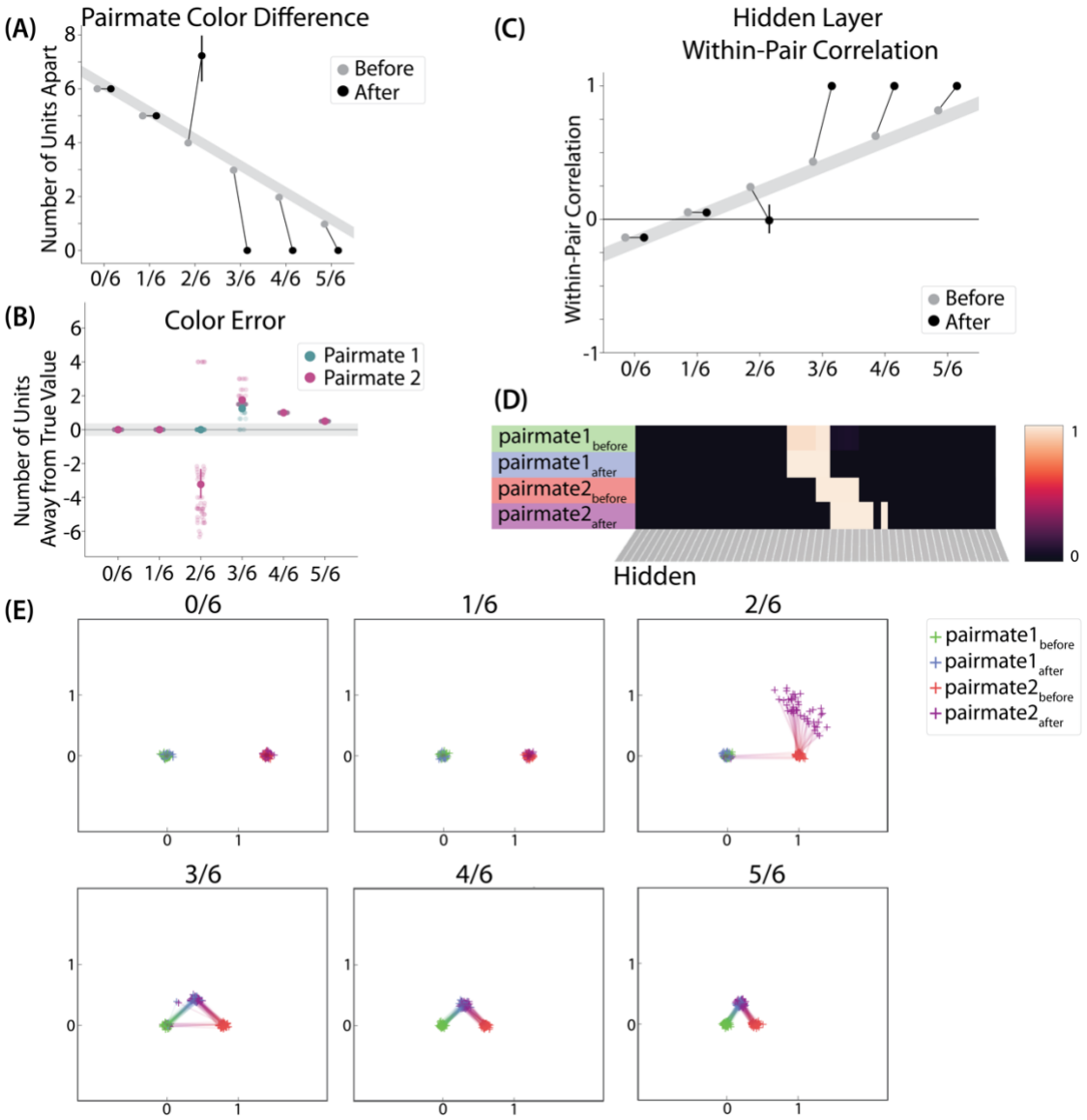


Figure 3.3: Model of Chanales et al. (2021), Results. (A) *The distance between the center of mass of output activity for A and B is used to measure whether repulsion or attraction occurs. Error bars for all figures indicate the 95% confidence interval. The grey bar indicates what color difference would be after learning (black dots) if no change happens—above is repulsion, and below is attraction. In the lower levels of overlap (0/6 and 1/6), color distance remains unchanged. In the medium level of overlap (2/6), repulsion occurs, shown by the increase in the number of units between A and B. In the higher levels of overlap (3/6, 4/6, and 5/6), attraction occurs, shown by the decrease in the number of units between A and B.* ***(B)*** *Here, color error (defined as the output layer distance between the "guess" center of mass and the "correct" center of mass) is shown for each pairmate (negative values indicate repulsion). When repulsion occurs (2/6), the change is driven by a distortion of pairmate 2.* ***(C)*** *Pairmate similarity is measured by the correlation of the hidden layer patterns both before and after learning. Here, above the grey line indicates integration, and below indicates repulsion. The within-pair correlation decreases (differentiation) when competitor overlap is moderate (2/6). Within-pair correlation increases (integration) when competitor overlap is high (3/6, 4/6, and 5/6).* ***(D)*** *Four activity patterns of a sample run in the 2/6 condition are shown, aligned with the corresponding units in the hidden layer: pairmate* 1_{before}*, pairmate* 1_{after}*, pairmate* 2_{before}*, and pairmate* 2_{after}*. Brighter colors indicate the unit is more active. In this run, pairmate A (left) is pairmate 1, so pairmate B (right) distorts away from pairmate A.* ***(E)*** *MDS plots for each condition are shown, to illustrate the pattern of representational change in the hidden layer. The same four patterns as in* ***Fig 3.3D*** *are plotted for each run. A jitter was applied to all points, and distance between pairmate* 1_{before} *and pairmate* 2_{before} *was adjusted to reflect increasing baseline similarity. Asymmetry in distortion can be seen in 2/6 by the movement of pairmate* 2_{after} *away from pairmate 1. In conditions that integrate, most runs lead to symmetric distortion, although some runs in the 3/6 condition lead to asymmetric integration, where pairmate 2 moves toward pairmate* 1_{before}*.*

layer is meaningful (with similar units representing similar colors), we could measure attraction and repulsion by the change in the distance between the centers of mass (**Fig 3.3A**): If A and B undergo attraction after learning, the distance between the color centers of mass would decrease. If A and B undergo repulsion, the distance would increase.

We found that for the 0/6 and 1/6 conditions (which correspond to the lower color similarity conditions of 72˚ and 48˚, respectively) there is no difference in the color center of mass before and after training. As similarity is increased to the 2/6 condition (which corresponds to the 24˚ similarity condition), the distance between the centers of mass increases after learning, indicating repulsion. When similarity is increased further for the 3/6, 4/6, and 5/6 conditions, the distance between the centers of mass decreases with learning, indicating attraction. Note that in the highest similarity condition (5/6), the initial distance between the centers of mass is only one unit, so the amount of possible attraction is very small. This may explain why the highest color similarity condition tested in Chanales et al. (2021) [102] (6˚) led to no change in color error.

This version of our model replicates the findings of Chanales et al. (2021) [102]. That is, if color similarity is increased to a high enough level, this can lead to a repulsion effect in recalled color values. If the similarity is increased even further, the repulsion effect is eliminated.

Differentiation and integration

Chanales et al. (2021) [102] is a behavioral study, and thus has no neural measure of differentiation or integration in the brain. Nonetheless, the attraction and repulsion effects presumably arise from neural changes. Moreover, a study based on the Chanales et al. (2021) [102] paradigm found that the level of distortion of the color memories is predicted by the amount of neural differentiation for those pairmates in parietal cortex [103]. We, therefore,

wanted to investigate how representational change manifests in our model's hidden layer. We expected that behavioral repulsion and attraction correspond to differentiation and integration in the hidden layer, respectively.

To measure this, we computed the Pearson correlation of the hidden layer's representation for A and B during the test epochs. If no representational change occurred, then this within-pair correlation would not change over time. In contrast, differentiation (or integration) would result in a decrease (or increase) in within-pair correlation over time.

The results of our model indicate that representational change happens in the same pattern as the repulsion/attraction results; the lower overlap conditions (i.e., 0/6 and 1/6) lead to no change in the hidden layer, moderate overlap (i.e., 2/6) leads to a decrease in within-pair correlation (differentiation), and higher overlap (3/6, 4/6, and 5/6) lead to an increase in within-pair correlation (integration) (**Fig 3.3C**).

Just as we observed when looking at attraction and repulsion effects, the amount of integration (i.e. the change in correlation before and after training) in the highest overlap condition (5/6) is small, but this is because pairmates in the 5/6 condition already have such similar hidden layer representations, that there is a limit on how much integration could occur. In fact, all three "high" overlap conditions have correlation values approaching 1 after training.

Asymmetry of representational change

The analyses above look at the relative similarity of the representations of A and B as a whole, replicating the experimental results of Chanales et al. (2021) [102]. In addition, our model allows us to track the change in the representations of the two stimuli individually, by looking at the change in the units belonging to each pairmate over time. This revealed that the

representational change is often asymmetric; for all of the conditions that led to differentiation/repulsion, and some of the runs that led to attraction/integration, the order of presentation determines which item distorts and which anchors.

For the output layer, we tracked how the center of mass for both pairmates changes over time (**Fig 3.3B**). For each pairmate, we calculated the difference between the center of mass at the end of learning compared to the true center of mass. A distortion away from the competitor is coded as negative, and toward the competitor is positive. When repulsion occurred in the 2/6 condition, the item shown first (pairmate 1) "anchors" in place (i.e., the final color report is unchanged and accurate). Repulsion arises from pairmate 2 moving away from its competitor.

This asymmetry was also observed in the hidden layer (**Fig 3.3D** and **3.3E**). We used multidimensional scaling (MDS) to visualize how the hidden layer representations changed over time. Given patterns of hidden layer activity, MDS represents them as points in a 2D plot so that the distances among each pair of points correspond to the dissimilarity between those samples. We made an MDS plot for four patterns for each run: the initial hidden layer pattern of the item shown first ("pairmate 1_{before}"), the initial hidden layer pattern of the item shown second ("pairmate 2_{before}"), the final hidden layer pattern of the item shown first ("pairmate 1_{after}"), and the final hidden layer pattern of the item shown second ("pairmate 2_{after}"). Since the 0/6 and 1/6 conditions do not give rise to representational change, the MDS plot unsurprisingly shows the before and after patterns for both pairmates remain unchanged. For the 2/6 condition (which shows differentiation), the pairmate 1_{before} item remains clustered around pairmate 1_{after}. However, pairmate 2_{after} distorts, becoming relatively more dissimilar to pairmate 1, indicating that the differentiation is driven by pairmate 2 distorting away from its competitor. This asymmetry effect can also occur in the higher overlap conditions that lead to integration. The

explanation for whether the integration that occurs is symmetric or asymmetric is described in the following section.

Why does this asymmetry occur during differentiation? When competition first happens, the target claims the shared units and the competitor units loses its connection to shared units, resulting in a new state where the pairmates are no longer in competition (see **Video 2**). Instead, the target is reinforced, and the competitor is left in a vulnerable state that is ready to change.

In our model, this asymmetry manifests in a clear order effect—pairmate 1 anchors and pairmate 2 distorts. It is important to remember, however, that our model was built to target the first moment the competitor pops up adequately, not the first moment of any task. Consequently, while the order of presentation is important for the results of this model, it may not be manifested as an order effect in empirical studies (see "Discussion" for more).

Two different kinds of integration

One pattern we noticed was that integration could take two different forms, symmetric and asymmetric (**Fig 3.3B** and **3.3E**). In this version of the model, the symmetric form of integration is more common. This can be seen in the MDS plots for 3/6, 4/6, and 5/6; both pairmate 1_{after} and pairmate 2_{after} mutually move toward each other. This is because both ended up connecting to all the units that were previously connected to either individually. However, in the 3/6 condition, some runs lead to an asymmetric form of integration where pairmate 2 distorts toward pairmate 1. (**Fig 3.3E**).

Why does this difference happen? The difference depends on the relative balance of competitor-to-competitor connections compared to competitor-to-shared connections after the first trial (**Fig 3.4**). Generally, competitor-to-competitor coactivity is less than competitor-to-shared coactivity, which is less than target-to-shared coactivity. In the 2/6 condition (**Fig 3.4A**), competitor-to-competitor coactivities remain on the left side of the U-shaped function, and are left unchanged. In the 4/6 and 5/6 conditions, since the competitor activity is so high, competitor-to-competitor coactivities fall on the right side of the U-shaped function (**Fig 3.4B**). It follows, then, that there is some overlap amount where the competitor-to-competitor connections falls in the dip of the U-shaped function. This is what happens in some runs in the 3/6 condition: On trial 1, some competitor-to-competitor connections sever. If competitor-to-shared connections strengthen at the same time, the internal instability of pairmate 2 leads to a flip where pairmate 2 moves toward pairmate 1 (**Video 4**).

Thus, two kinds of integration can occur—one where both items pick up all units that initially belonged to either (symmetric), and one where pairmate 2 moves toward pairmate 1 (asymmetric). Generally, as competitor

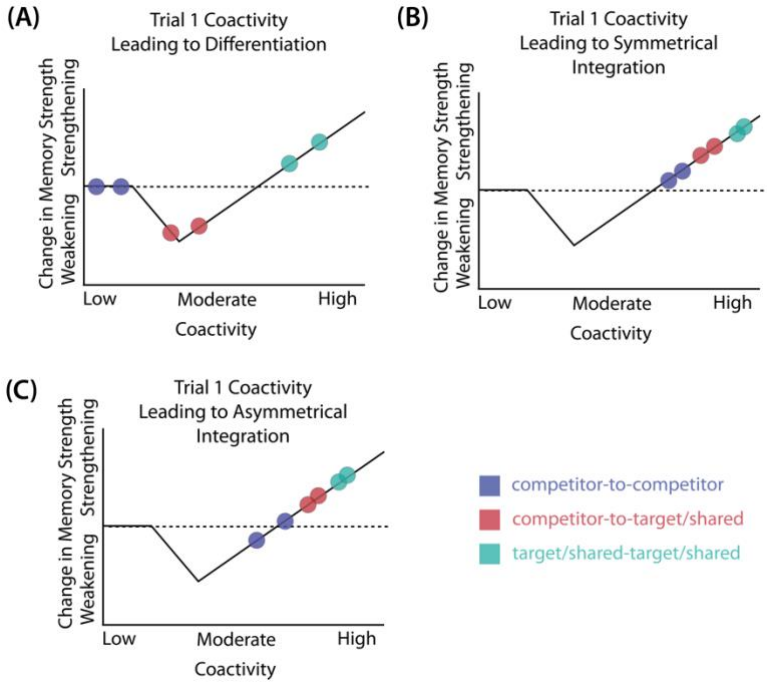


Figure 3.4: Schematic of coactivities on Trial 1. *Generally, competitor-to-competitor coactivity is less than competitor-to-shared coactivity, which is less than target-to-target coactivity.* ***(A)*** *This is a schematic of the typical arrangement of coactivity in the hidden-to-hidden connections and item-to-hidden connections in the 2/6 condition. Since the competitor-to-shared connections fall in the dip of the U-shaped function, they sever, which leads to differentiation.* ***(B)*** *This is the typical arrangement of coactivity in the higher overlap conditions, like 4/6 and 5/6. Here, all connection types are on the right side of the U-shaped function, leading to integration. Specifically, all the units connect to each other more strongly, leading to all units previously associated with either pairmate A or B joining.* ***(C)*** *As competitor pop up increases to move from the situation in (A) to (B), there is a level of competitor overlap where the competitor-to-competitor coactivity falls in the dip of the U-shaped function. When enough competitor-to-competitor connections weaken while the competitor-to-shared connections strengthen, this leads to an imbalance resulting in an asymmetric integration. This happens in some runs of the 3/6 condition.*

activity is raised, competitor-to-competitor coactivity is raised, and it is more likely the integration becomes symmetric.

Differentiation requires a high learning rate but integration does not

We found that incremental learning does not lead to differentiation. In our model, the change in weighted connection strengths on each trial is multiplied by a learning rate ("LRate"), which is usually set to be 1. Lowering the LRate value, consequently, leads to smaller learning increments. When we cycle through LRate values for all projections other than the output-to-output connection (where LRate is zero), we find that differentiation fails to occur in the 2/6 condition if LRate is too low (**Fig 3.5A**). This is because if the competitor-to-shared connections are not fully severed on Trial 1, the (formerly) shared units may still receive enough excitation to activate on Trial 2, and there is no pressure for new units to join the representation (see Video 5). The pattern of representational change is not affected by LRate in the low overlap conditions (0/6 and 1/6) or high overlap conditions (3/6, 4/6, and 5/6).

The failure to differentiate with lower learning rates can take two forms: Either no change occurs, or sometimes integration occurs (**Fig 3.5B**). As shown in the video above, the integration that may result is the same asymmetric integration that occurs sometimes in the 3/6 condition. It arises because the shared-to-competitor connections have somewhat weakened, while the shared-to-target connections are still high. If the shared units then become active on Trial 2, it leads to an imbalance that may create a "swap."

A consequence of the requirement of a high learning rate for differentiation is that differentiation has to occur abruptly if it does occur at all. Essentially, after competitor-to-shared connections are weakened, one of two outcomes occur. If the learning rate is high enough to

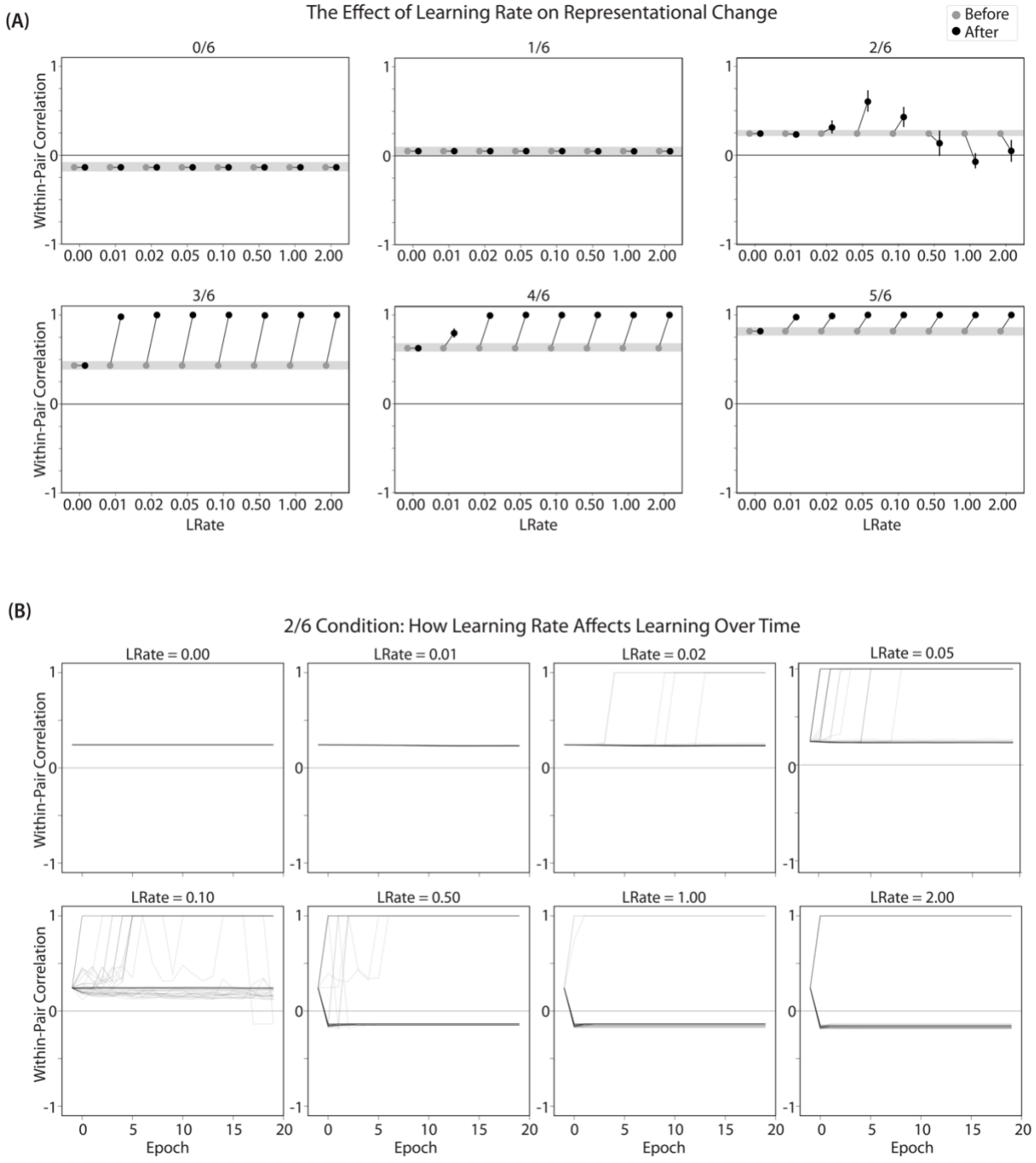


Figure 3.5: Learning rate and representational change. (A) *The learning rate (LRate) parameter was adjusted and within pair correlation of the hidden layer for each overlap condition was calculated. The default LRate for simulations in this paper is 1. In the low overlap conditions (0/6 and 1/6), adjusting the LRate has no impact on representational change. In the 2/6 condition, differentiation does not occur if the LRate is lowered. In the high overlap conditions (3/6, 4/6, and 5/6), adjusting the LRate has no impact on representational change.* ***(B)****. For each LRate value tested, the within-pair correlation over time in the 2/6 condition is shown, where each line is a separate run. When differentiation does occur, it is abrupt and total.*

effectively sever the competitor-to-shared connections, differentiation will occur on the subsequent trial. If the severing is not sufficient, the former shared units will be reincorporated and no differentiation will occur. This can be seen by the abrupt differentiation in **Fig 3.5B**. This is also supported by empirical studies—for instance, Wanjia et al. (2021) [109] showed that behavioral expressions of successful learning are coupled with a temporally abrupt, stimulus-specific decorrelation of CA3/dentate gyrus activity patterns for highly similar memories.

A useful analogy may be escape velocity from astrophysics. Spaceships need to be going a certain speed to escape the pull of gravity. Similarly, the competitor representation needs to get a certain distance away from its pairmate in one trial, or else it will get "pulled" back in.

Important lessons

Our model of Chanales et al. (2021) [102] shows that the NMPH can explain the results observed in the study—moderate color similarity can lead to repulsion, and if the color similarity is increased beyond that point, that repulsion is eliminated. Our model shows how the behavioral changes in this paradigm are linked to differentiation and integration of the underlying neural representations. The simulations also enrich our NMPH account of these phenomena in several ways, beyond the verbal account provided in Chapter 1. In particular, the simulations expose some important boundary conditions for *when representational change can occur* according to the NMPH (showing that differentiation depends on a large learning rate, but integration does not), and the simulations provide a more nuanced account of *exactly how*

representations change (showing that differentiation driven by the NMPH is always asymmetric, whereas integration is sometimes asymmetric and sometimes symmetric).

There are several aspects to Chanales et al. (2021) [102] that were left out of our model (for instance, they additionally used an interleaved task where the colored object appeared as a cue and the participants had to select the correct face), but our goal was to show how the simplest form of their paradigm could lead to the distortion effects observed. Future simulations can assess whether these other experiment details affect the predictions of the model.

Model of Favila et al. (2016): Similar vs different predictive associations

Key experimental findings

Favila et al. (2016) [10] provided neural evidence for differentiation following competition, using a shared associate to induce competition between pairmates. In this study (**Fig 3.6A**), participants were instructed to learn scene-face associations. Later, during the repeated face-test task, participants were shown a scene and asked to pick the correct face from a bank of faces. Scenes were made up of highly similar pairs (e.g., two bridges, two barns). Sometimes, two paired scenes predicted the same face, sometimes different faces, and sometimes no face at all (in the latter case, the paired scenes appeared in the study task and not the face-test task). The participants were never explicitly told that some scenes shared a common face associate.

When the two scenes predicted different faces, the hippocampal representations for each scene were relatively orthogonalized—two pairmate scenes were just as neurally overlapping as they were to non-pairmate scenes. However, when the two scenes predicted the same face,

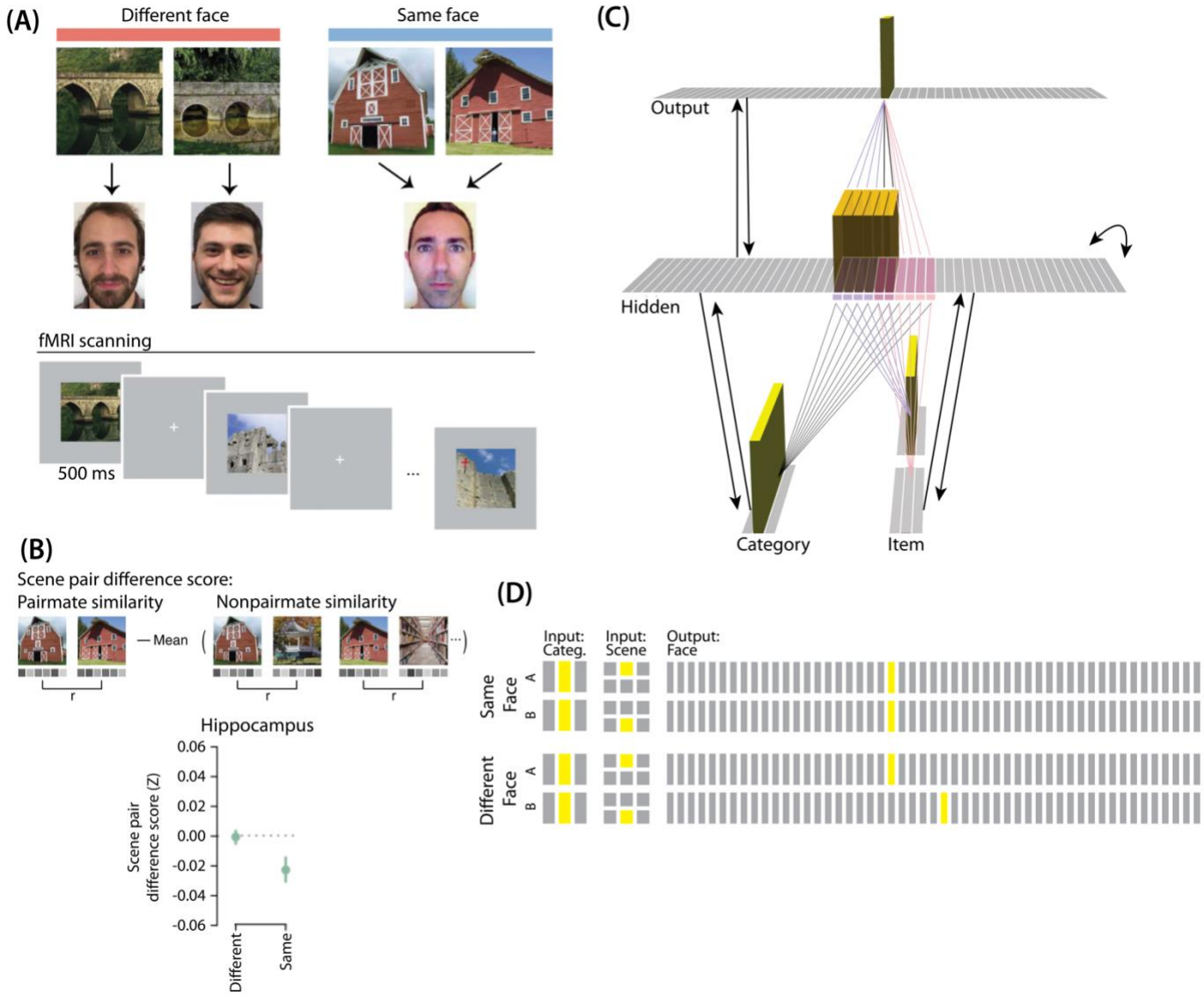


Figure 3.6: Modeling Favila et al. (2016). (A) *Participants in Favila et al. (2016) learned to associate individual scenes with faces. Scenes were actually pairs of different categories (e.g., barns), and pairmates could be associated with the same face, different faces, or no face at all (not shown). Subjects were scanned while looking at each individual scene in order to get a measure of neural representations for each scene.* ***(B)*** *Neural similarity was measured by a correlation of scene patterns. A scene pair difference score was calculated by subtracting non-pairmate similarity from pairmate similarity; this measure shows the relative representational distance of pairmates. Results for the different face and same face condition in the hippocampus are shown: linking scenes to the same face led to a negative scene pair difference score, indicating that scenes became less similar to each other than they were to non-pairmates (differentiation).* ***(C)*** *To model this study, we used the same basic structure as described in "Basic Network Properties." In this model, the category layer represents the type of scene (e.g., barn, bridge, etc), and the item layer represents the individual scene. The output layer represents the face associate. Activity shown is for pairmate A in the same face condition. Category-to-hidden, item-to-hidden and hidden-to-hidden connections are pre-wired similarly to the 2/6 condition of our model of Chanales et al. (2021) (see* ***Fig 3.1****). Additionally, the hidden A and B units have random, low level connections to all output units, but are additionally pre-wired to connect strongly to either one or two units in the output layer—in the different face condition, hidden A and B units are pre-wired to connect to two different face units, but in the same face condition, they are pre-wired to connect to the same face unit.* ***(D)*** *This model has two conditions—same face and different face. The only difference between conditions is whether the hidden A and B units connect to the same or different face unit in the output layer.*

differentiation resulted, and the two similar scenes were less similar to each other than they were to non-pairmate scenes (**Fig 3.6B**).

Potential NMPH explanation

Crucially, these results are the opposite of what would be predicted by error-driven learning, which predicts that pairing two stimuli with the same outcome would lead to integration, not differentiation. The NMPH, however, could potentially explain these results: The linking of the items to the shared face associate provides an additional pathway for excitation to spread to the competitor. This may increase the activity of the competitor to moderate levels, pushing competitor-to-target coactivity into the dip of the U-shaped function, leading to differentiation. Activation of the competitor in the "different" face condition would be lower, and so there is less competition to drive representational change.

Model set up

Model architecture

Our model of Chanales et al. (2021) [102] can be adapted for Favila et al. (2016) [10]. Instead of mapping from an object and face to a color, as in Chanales et al. (2021) [102], the Favila et al. (2016) [10] study involves learning mappings between scenes and faces. Also, the way in which competition is manipulated is different across the studies: In Chanales et al. (2021) [102], competition is manipulated by varying the similarity of the stimuli (specifically, the color difference of the objects), whereas in Favila et al. (2016) [10], competition is manipulated by varying whether pairmate scenes are linked to the same vs different face.

To adapt our model for Favila et al. (2016) [10], the interpretation of each layer must be altered to fit the new paradigm (**Fig 3.6C**). The category layer now represents the category for the scene pairmates (i.e., barn, bridge, etc.). The item layer represents the individual scene (i.e., where A represents $barn_1$ and B represents $barn_2$). Just as before, any given stimulus includes an external input from both of these input layers—the category layer unit is consistent for A and B, but the item layer unit differs.

The output layer in this model represents the face associate, such that each individual output layer unit could be thought of as a single face. For this model, the ordering of units in the output layer is not meaningful—two units next two each other in the output layer are no more similar than to any other units.

Knowledge built into the network

As before, we were interested in the moment that competition first happens, so we pre-wired connections as if some initial learning had occurred. We pre-wired the connections from the hidden layer to both input layers (and hidden layer to itself) to be similar to the 2/6 condition of our model of Chanales et al. (2021) [102], so there is some baseline amount of overlap between A and B (i.e., the similar-looking scenes).

To mimic the learning of scene-face associates, all hidden A units are connected to a unit in the face layer, and all hidden B units are connected to a unit in the face layer. In the different face condition, A and B hidden units are connected to different face units, to reflect that the two scenes were predictive of two separate faces. In the same face condition, A and B hidden units connect to the same face unit, to reflect that A and B predict the same face.

Manipulation of competitor activity

Competitor activity is modulated by the similarity of the predictive consequences in this version of the model—whether the hidden units for pairmates A and B are pre-wired to connect strongly to the same unit or different units in the output layer. Stronger connections to the same face unit should provide an extra conduit for excitation to flow to the competitor units.

Task

The task performed by the model is the same as for our model of Chanales et al. (2021) [102]—we clamped the external input for the scenes and allowed activity to spread through the hidden layer to the output layer so it could make a guess for the correct face. No correct answer is shown, and no error-driven learning is used. While the exact values of parameters varied in this model to account for the different process of inhibitory competition in this study (see "Methods" for details), inhibition and oscillations were implemented in the same way as before.

Results

The results of our model for Favila et al. (2016) [10] aligned closely with the results we found for our model of Chanales et al. (2021) [102]. However, there is no repulsion and attraction in the output layer for this version of our model, because the output layer represents the face associate, which is a discrete item rather than a continuous behavioral measure. Instead, we measured the representational change in the hidden layer only.

Differentiation and integration

The different face condition in this model led to no representational change (see **Video 6**) whereas the same face condition led to differentiation (see **Video 7**), as measured using within-pair correlation (**Fig 3.7A**). In this version of the model, the topography of the hidden layer is not meaningful other than the units assigned to A and B, so the new units that are added to the representation could be on either side. Differentiation is indicated by the fact that new units are added that did not previously belong to either pairmate.

Asymmetry of representational change

As in other fMRI studies of representational change, this study defines differentiation or integration as a score that compares the neural similarity of pairmates versus non-pairmates [10,31,103,109]. This is a useful measure, but does not allow us to track the representational

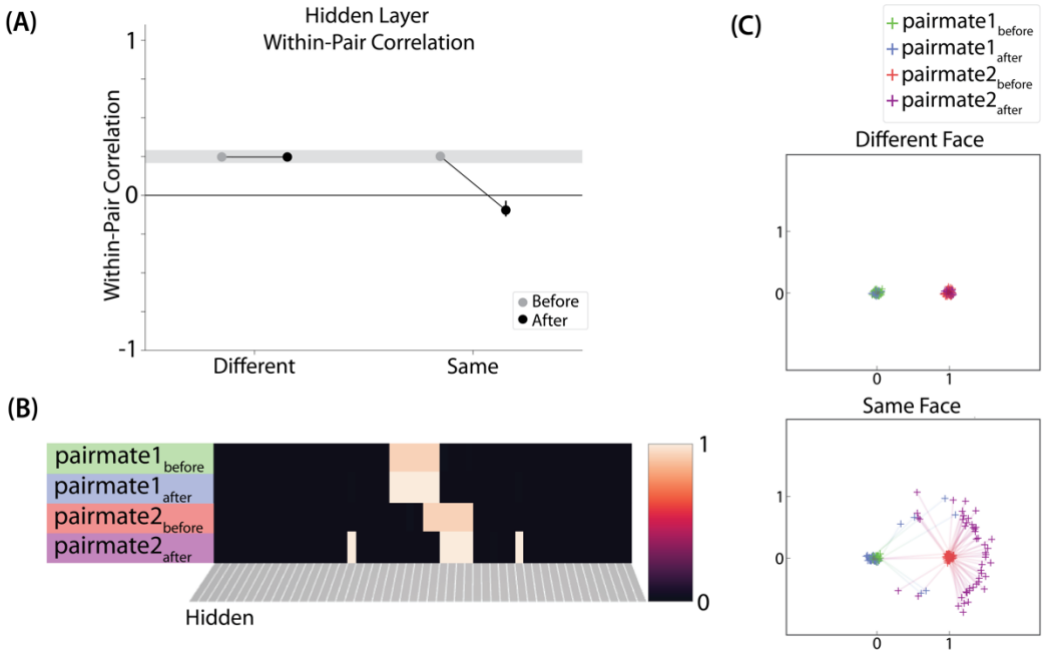


Figure 3.7: Model of Favila et al. (2016), Results. (A) *Within-pair correlation between A and B hidden layer representations before and after learning. In the same face condition, the within-pair correlation is reduced after learning, indicating differentiation.* ***(B)*** *Activity patterns of both pairmates in the hidden layer before and after learning for a sample same face run are shown. Asymmetry in distortion can be seen in how pairmate 2 picks up additional units that did not previously belong to either A or B.* ***(C)*** *MDS plots for each condition are shown to illustrate the pattern of representational change in the hidden layer. In the same face condition, asymmetric differentiation is shown in that the pairmate 2_{after} representation generally moves further away from pairmate 1 in representational space while pairmate 1 generally does not change.*

change in individual items, which we are able to do in our model by tracking the change in units associated with each stimulus. The representational change in the hidden layer shows the same kind of asymmetry that occurs in our model of Chanales et al. (2021) [102] (**Fig 3.7B**). Since the new added units could be on either side, this means that a change in the representation may not result in a shift in the center of mass of activity. But the representational change in this model is asymmetric as well, in that pairmate 1 consistently anchors whereas pairmate 2 changes, picking up units that did not belong to either representation initially. The MDS plot, therefore, is useful to see the structure of this asymmetric representational change (**Fig 3.7C**). In the different face condition, no representational change occurs. But in the same face condition, pairmate 2_{after} is consistently distorted away from pairmate 1 (note that, as shown in **Fig 3.7C**, some runs fail to differentiate and instead lead to integration—see the "Discussion" for how conditions that usually lead to differentiate may fail and lead to integration instead).

Important lessons

Our model shows that the NMPH can predict the results of Favila et al. (2016) [10], where learning about stimuli that share a paired associate can lead to differentiation. The model shows how linking items with the same or different associates can modulate competitor activity and—through this—modulate the representational change that occurs; as in our simulation of Chanales et al. (2021) [102], we found that the NMPH-mediated differentiation was asymmetric.

Model of Schlichting et al. (2015): Blocked vs interleaved learning

Key experimental findings

Another study with implications for the NMPH is Schlichting et al. (2015) [11]. This study examined how the learning curriculum affects representational change in different brain regions. In this study, participants learned to link novel objects with a common associate (i.e., AX and BX). Sometimes these associates were presented in a blocked fashion (all AX before any BX) and sometimes they were presented in an interleaved fashion (**Fig 3.8A**).

The analysis focused on the hippocampus, the medial prefrontal cortex (mPFC) and the inferior frontal gyrus (IFG). Some areas (e.g., right posterior hippocampus) showed differentiation for both blocked and interleaved conditions, some areas (e.g., left mPFC) showed integration for both conditions, and others (e.g., right anterior hippocampus) showed integration in the blocked condition but differentiation in the interleaved condition (**Fig 3.8B**).

Potential NMPH explanation

The NMPH can explain how the results differ by brain region, since the overall level of inhibition in an area can limit the competitor's activity. For instance, regions that tend to show differentiation (like posterior hippocampus) have sparser activity [69]. The higher inhibition in these areas could cause the activation of the competitor to fall into the moderate range (leading to differentiation). For regions where inhibition is lower, competitor activation may fall into the high range (leading to integration).

The result that some regions (e.g., right anterior hippocampus) show differentiation in the interleaved condition and integration in the blocked condition could also be explained by the NMPH. We would expect that by the first BX trial, the connections between A and X would have been strengthened much more in the blocked condition compared to the interleaved condition. That stronger connection could provide an additional pathway for activity to flow to the competitor in the blocked condition. Consequently, the activity in the blocked condition

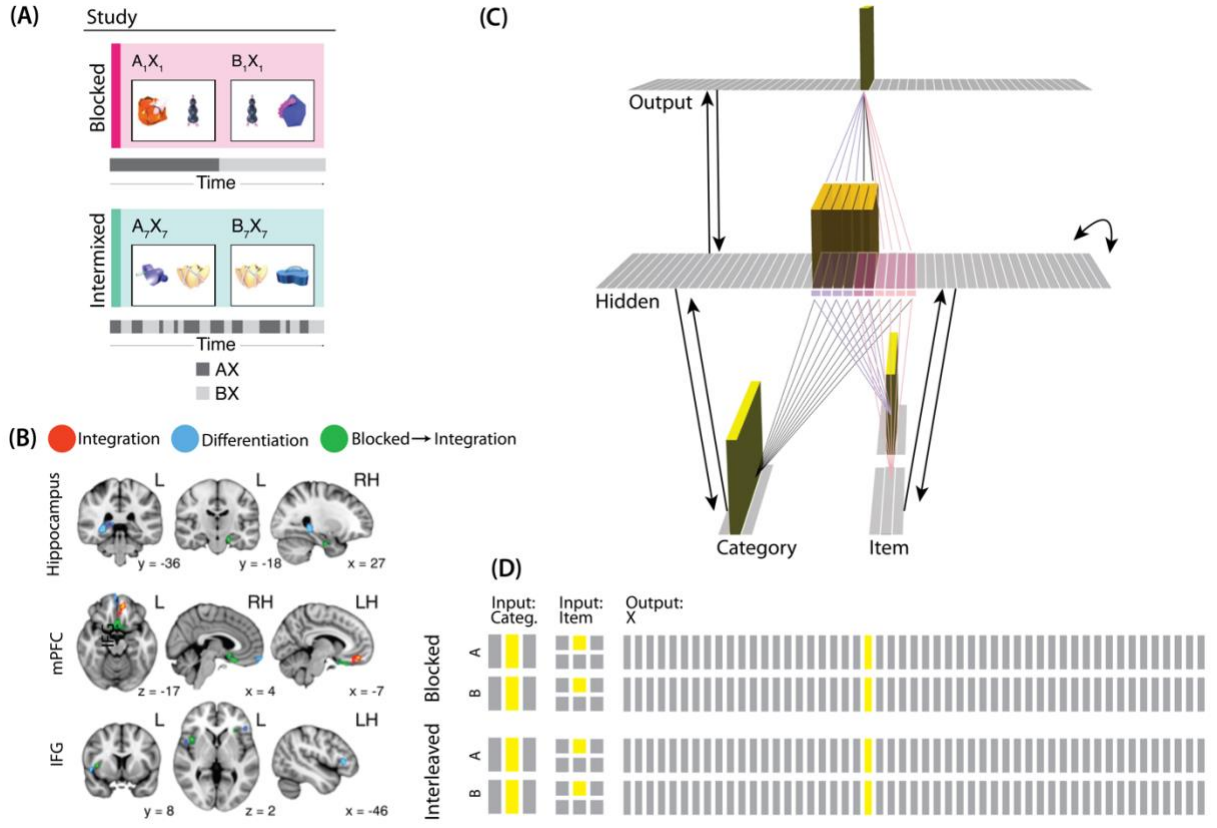


Figure 3.8: Modeling Schlichting et al. (2015). (A) *Participants in Schlichting et al. (2015) learned to link novel objects with common associate (i.e., AX and BX). Sometimes these associates were learned in a blocked design (i.e., all AX before all BX), and sometimes they were learned in an interleaved design. The items were shown side-by-side, and subjects were not explicitly told the structure of the shared items. Before and after learning, subjects were scanned while observing each item alone, in order to get a measure of the neural representation of each object.* ***(B)*** *The analysis found that some areas (e.g., right posterior hippocampus) showed differentiation for both blocked and interleaved conditions, some areas (e.g., left mPFC) showed integration for both conditions, and other areas (e.g., right anterior hippocampus) showed integration in the blocked condition but differentiation in the interleaved condition.* ***(C)*** *To model this study, we used the network structure described in "Basic Network Properties," with the following modifications: the structure of the network was similar to the same-face condition in our model of Favila et al. (2016) (see* ***Fig 3.6****), except we altered the connection strength from units in the hidden layer to the output unit X to simulate what would happen depending on the learning curriculum. In both conditions, we assume that some initial amount of BX learning has taken place (possibly on the current trial), and the pre-wired connection from hidden B units to the output X unit is set to .7. In the interleaved condition, the connection strength from hidden A units to the output X unit is .8, to reflect some amount of initial AX learning. In the blocked condition, the connection strength from hidden A units to the output X unit is raised to .999, to reflect extra AX learning.*

would fall further to the "right" of the U-shaped function compared to the interleaved condition, leading to a situation where integration could happen in the blocked condition but differentiation happens in the interleaved condition.

Model set up

Model architecture

In this version of the model, the category layer can be thought of as representing the kind of novel objects used as stimuli. In Schlichting et al. (2015) [11], A, B, and X are all individual items that were shown as stimuli in the same fashion, except for the fact that X was shown alongside A or B. We decided to represent A and B in the input layer and to represent X as a single unit in the output layer (note that, because connections in the model are bidirectional and symmetric in strength, the X "output" was able to influence activation elsewhere in the network; this is critical for our explanation of these findings). Putting A and B in the same layer, where their representations directly compete, captures the fact that A and B were mutually exclusive (i.e., they were never viewed together); putting X in a different layer made it easier for the model to represent X at the same time as A or B. This approach also makes the isomorphism to Favila et al. (2016) [10] more clear—just as a shared face associate in that study served as a conduit for activation to spread to the scene pairmate, the shared "X" object associate in Schlichting et al. (2015) [11] serves as a conduit for activation to spread between B and A.

Knowledge built into the network

The connections between layers for this model are similar to the connections we used in the same face condition in our model of Favila et al. (2016) [10]. Since both A and B have a shared associate (X), all hidden A and B units are connected to the same unit in the output layer. However, we adjusted the connection strength to the shared X unit in order to simulate blocked or interleaved learning (see **Fig 3.8C** caption for details).

Modulation of competitor activity

We hypothesized that competition would be modulated by the blocked vs interleaved manipulation. If AX and BX are learned in a blocked design, then the connections between A (in the hidden layer) and X (in the output layer) will be stronger by the first BX trial than in the interleaved condition. That extra strength can provide a pathway for more excitation to flow from X to during BX trials. In our simulations, we operationalized this difference by varying the strength of the connection between A units (in the hidden layer) and X units (in the output layer; see **Fig 3.8** caption for details).

Task

The task is generally the same as that for the other versions of our model. External input for A or B (given by the two input layers) is clamped, and activity is allowed to flow through the hidden layer to the output layer, producing a guess. This deviates somewhat from the task in the original study, which showed AX or BX side-by-side. We could have clamped the X external values so it would be shown rather than guessed, but we decided to have the activity flow through the network in this way so it would be consistent with the other versions of our model.

Nonetheless, this change does not have any impact on the outcome of the simulation—X ends up being strongly active here even though it is not being externally clamped, so the dynamics of activation elsewhere in the network (and the resulting learning) end up effectively looking the same as if external clamping had been applied.

In the blocked condition, we assume AX has been fully learned, so we start with a BX training trial and no AX trials follow. To be consistent, we also start the interleaved condition with a BX training trial; however, in the interleaved condition, both AX and BX may follow after that initial BX trial.

While the exact values of parameters varied in this model to account for the different form of inhibitory competition (see "Methods" for details), inhibition and oscillations were implemented in the same way as before.

Modulation of inhibitory dynamics

For this version of our model, we added one extra manipulation. To model the varying differentiation and integration results in different brain regions, we varied the amount of inhibitory oscillations in the hidden layer. If the amplitude of the sinusoidal function on inhibition is bigger (or smaller), then the competitor is allowed to pop up more (or less), affecting the outcome of learning. We will mostly discuss the version of the model where the amplitude is set to be .0623, but we additionally allowed less competitor pop up (by lowering the oscillation amplitude to .0535) and more competitor pop up (by raising the oscillation amplitude to .09) to simulate representational change in brain areas that have more or less restrictions on competitor activity.

Results

Differentiation and integration

The results of our model of Schlichting et al. (2015) [11] followed a similar pattern to our model of the other experiments. Examining the correlation between A and B hidden layer representations, we found that the interleaved condition leads to differentiation (see **Video 8**) whereas the blocked condition (where the AX connection weight is stronger) leads to integration (see **Video 9**) (**Fig 3.9A**).

Asymmetry of representational change

The results again indicate that differentiation is asymmetric. Just like our model for Favila et al. (2016) [10], the topography of the hidden layer is not meaningful other than the units for A and B, so the units that get added could be anywhere else on the layer. The asymmetric differentiation can be seen in the interleaved condition by how pairmate 2 picks up units that did not previously belong to either representation (**Fig 3.9B**). The MDS plot shows how this asymmetric distortion happens systematically (**Fig 3.9C**). In the interleaved condition, it is pairmate 2 that consistently distorts away from pairmate 1.

The blocked condition shows integration that is mostly symmetric, but there are many runs that show asymmetric integration (just like in the 3/6 condition of our model of Chanales et al. (2021) [102]). The reason for the asymmetric integration is the same here—sometimes not all the competitor-to-competitor connections reach the right side of the U-shaped function, and they sever. When these connections sever while the competitor-to-shared connections strengthen, that imbalance can cause a flip from pairmate 2 toward pairmate 1.

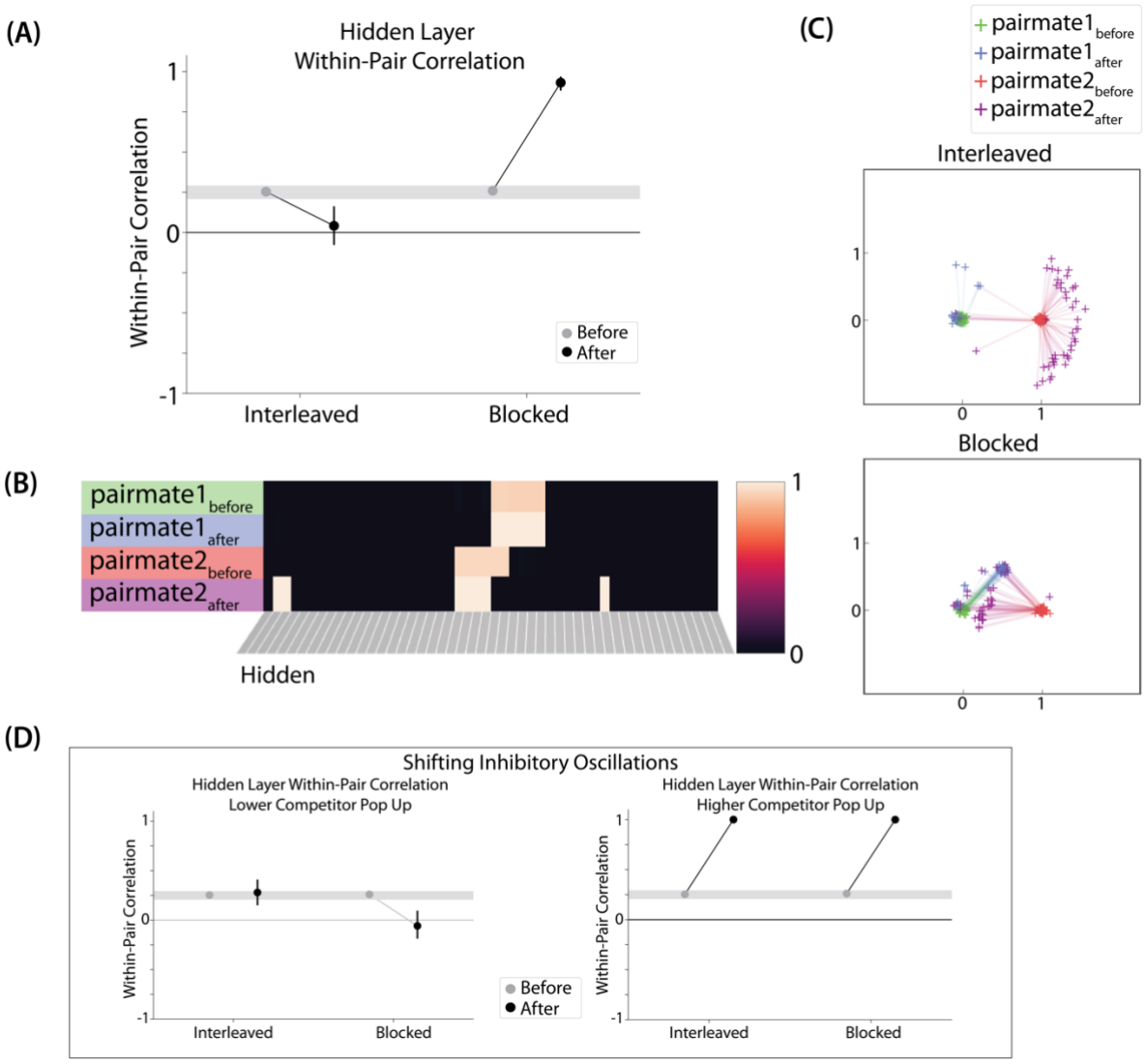


Figure 3.9: Model of Schlichting et al. (2015), Results. *Note that in this version of our model, pairmate B is always the stimulus shown first. We label pairmate A as pairmate 2 in these plots, even though in the blocked condition, there are no subsequent AX trials (AX is only shown at test).* ***(A)*** *Within-pair correlation between hidden layer A and B representations is shown before and after learning. In the interleaved condition, the within-pair correlation is reduced after learning, indicating differentiation. In the blocked condition, the within-pair correlation increases, indicating integration.* ***(B)*** *Activity patterns of both pairmates in the hidden layer before and after learning are shown for a sample run. Asymmetry in distortion can be seen in how pairmate 2, but not pairmate 1, picks up additional units that did not previously belong to either representation.* ***(C)*** *MDS plots for each condition are shown to illustrate the pattern of representational change in the hidden layer. In the blocked condition, the pairmates integrate and move toward each other. This integration is mostly symmetric, but on many trials it is asymmetric, shown in how pairmate 2 moves toward pairmate 1 rather than pairmates 1 and 2 meeting in the middle. In the interleaved condition, asymmetric differentiation occurs, in that pairmate 2 moves away from pairmate 1.* ***(D)*** *In this model, we additionally adjusted the oscillation amplitude to change the level of competitor pop up, to investigate how these results could be altered in different brain regions with different inhibitory dynamics. No parameters other than inhibitory oscillation level were changed. When the oscillation strength is reduced from .0623 to .0535, less competitor pop up happens, and the blocked, but not interleaved, condition leads to differentiation. When the oscillation amount is raised to .09, more competitor pop up happens, and both conditions lead to integration.*

Adjusting oscillation amount changes representational change

In this model, we additionally adjusted the amplitude of oscillations, to simulate how inhibitory dynamics in different brain regions may affect these results (**Fig 3.9D**). When the competitor is less able to activate (as a result of lower inhibitory oscillations), only the blocked condition leads to differentiation. When the competitor is allowed to activate more (due to higher inhibitory oscillations), both conditions lead to integration. Thus, the NMPH can account for how—within a single experimental condition—brain areas with high levels of inhibition can show differentiation, and brain areas with low levels of inhibition can show integration.

Important lessons

The results of this version of our model show that the NMPH can predict the results of Schlichting et al. (2015) [11], where the learning paradigm (blocked or interleaved) was shown to affect representational change. Additionally, the model shows how the inhibitory dynamics of different brain regions can also affect these outcomes.

Discussion

Recent studies have presented a challenge to supervised learning models, showing that memories for similar stimuli can result in differentiation rather than integration. The goal of the current research was to provide a computational framework for an unsupervised learning account of representational change, and to show how a neural network model could achieve either differentiation or integration depending on the level of competitor activation. To do this, we built a neural network model for three studies, each of which modulated the amount of competitor activity in different ways: Chanales et al. (2021) [102] modulated competitor activation with

stimulus similarity; Favila et al. (2016) [10], modulated competitor activation via predictive consequences; and Schlichting et al. (2015) [11], modulated competitor activation by organizing learning in either a blocked or interleaved order. The Schlichting et al. (2015) [11] model also explored the effects of (simulated) regional differences in the strength of local inhibition, and the Chanales et al. (2021) [102] model additionally explored the effect of varying learning rate.

Our model provides an "existence proof" that a network imbued with NMPH learning can explain these findings. Using unsupervised, NMPH learning alone, we show how 1) increasing the similarity of color memories leads progressively to a repulsion effect and then an attraction effect; 2) pairing two stimuli with the same associate could lead to differentiation; and 3) learning in an interleaved vs blocked fashion leads to differentiation and integration, respectively, and that changing inhibitory dynamics affects these outcomes. Furthermore, in addition to replicating the results from the studies we simulated, the model provides several novel predictions: differentiation requires a rapid learning rate and is systematically asymmetric.

Differentiation requires a rapid learning rate and is sensitive to activation dynamics

Our model predicts that a high learning rate is required for differentiation, which means that differentiation must happen swiftly, if it is to happen at all: As shown in the results of the model for Chanales et al. (2021) [102], if the severing of connections between the unique features of the competitor and the (formerly) shared features is not complete after one trial, differentiation is reversed on the subsequent trial when the competitor activates.

We also found that differentiation is highly sensitive to activation dynamics. Anything that affects how strongly the competitor comes to mind can impact the kinds of representational change that are observed. We used a variety of methods to influence how strongly competitors

come to mind in the three models, but in particular, this is shown in the simulation where we modulated the oscillation amplitude. These results show that within one condition (interleaved or blocked), the outcome of representational change can switch depending on how much the competitor comes to mind.

These results have implications for where to look for differentiation in the brain. The results from our simulation where we adjusted the learning rate (the model for Chanales et al. (2021) [102]) suggest that differentiation will be most evident in regions with a high learning rate (for instance, the hippocampus). Additionally, the simulations where we adjusted oscillation amount (the model of Schlichting et al. (2015) [11]) suggests that differentiation will be most evident in regions where it is relatively hard to activate competitors, making it less likely that plasticity will shift from differentiation to integration. Taken together, these results suggest that within the hippocampus, differentiation will be most evident in regions with sparser activity (like dentate gyrus, and to a lesser extent, CA3). This is consistent with research showing these areas are biased toward differentiation; Molitor et al. (2021) [110] showed that strong memory reactivation led to differentiation in dentate gyrus and CA2/3, and Wammes et al. (2022) [111] found that learning of increasingly similar items triggered a U-shaped outcome for neural similarity in the dentate gyrus, such that moderately similar images differentiated in that region.

These results also have implications for how to run experiments to study differentiation. The high learning rate means that differentiation happens swiftly when it does occur—immediately after the competitor first comes to mind moderately. Consequently, our model predicts that representational change should occur rapidly in experimental studies. Even though our model is built so that the competition happens on the first trial, in reality it may take some learning of the stimuli before the pairmates are put in direct competition. Consequently, the key

moment of competition that leads to the representational change may come at any point over the course of learning. Additionally, the timing of that key moment may be different for each pair of stimuli. For instance, attentional lapses may lead some pairs to be missed at the start, leading them to come into competition later on. Or, variances in stimulus similarity may change the amount of learning required to put the items into competition.

If each pair of items has an abrupt change at a different point in the experiment, it could look as though learning is gradual if the time courses of all pairs are averaged during analysis. For instance, the authors of Chanales et al. (2021) [102] found that the percentage of responses away from the competitor increased gradually over learning blocks. Our model suggests that the repulsion effect may be abrupt if analyzed on an individual item basis.

It may be important, then, not only to track representational change over time, but to pinpoint the moment when this representational change happens for individual trials. Wanjia et al. (2021) [109] did this by associating neural and behavioral measures—they found that an abrupt decorrelation of highly similar memories in the brain is temporally coupled with behavioral expressions of learning.

Lastly, as we have previously argued, our model suggests that sweeping through parameters that continuously modulate competitor activation is a good way to find differentiation (in principle, increasing competitor activation should lead to no change, then differentiation, then integration). However, the sensitivity of differentiation in our model to activation dynamics indicates that, in certain circumstances, this is not guaranteed to work. For instance, if the competitor representation is damaged and cannot be re-accessed (which happens in some versions of the low learning rate simulation of Chanales et al. (2021) [102]) the model will "skip over" differentiation and show asymmetric integration instead.

Asymmetry of representational change

Our model shows that representational change is often asymmetric. In all versions of our model, when differentiation occurs, its effects are asymmetric, such that the item that first pops up as a competitor is the one that distorts. Conditions that lead to integration are sometimes symmetric and sometimes asymmetric (where the hidden layer representation for one item flips to the units for the other).

Our model then predicts that such an asymmetry should occur in experimental studies of representational change too. A common method of investigating representational change is some kind of pairmate similarity score [10,31,103,109], which subtracts the correlation of non-pairmate patterns from the correlation of pairmate patterns. Thus, a positive score indicates that two pairmates are more similar to each other than baseline, whereas a negative score indicates that two pairmates are less similar to each other than baseline. This is an important and useful measure to show differentiation. However, this measure is insufficient for investigating the asymmetry effects, since the measure does not distinguish which pairmate differentiates.

One way to investigate this would be to look at how the individual pairmate changes its representation over time. This may be more difficult to track, especially because there may not be a clear indication from the start which item will anchor and which will distort; our model predicts that it is the pairmate that pops up first as the competitor that will distort, but it is not trivial to identify when that happens in practice. For example, in our simulation of Chanales et al. (2021) [102], we simplified things by pre-wiring strong memories for both pairmates before presenting one of the pairmates (pairmate 1) to the network. In this simplified situation, pairmate 2 pops up as a competitor when pairmate 1 is studied, so pairmate 2 repels away from pairmate 1 and not vice-versa (**Fig 3.3B**). In the actual Chanales et al. (2021) [102] experiments, encoding

strength can vary across trials for numerous reasons (e.g., fluctuating attention); if pairmate 1 is the first to establish a strong representation, it will pop up and distort when pairmate 2 is next studied, but if pairmate 2 is the first to establish a strong representation, it will pop up and distort when pairmate 1 is next studied. This encoding variability makes it difficult to predict *a priori* which item will distort in the actual Chanales et al. (2021) [102] study.

One way to predict the asymmetry is to use a paradigm where we can be more confident which item will be the first to pop up as a competitor. For instance, an AX-BX paradigm using blocked trials as in Schlichting et al. (2015) [11] may be useful for this. In this design, all AX trials occur before any BX trials. So, it may be more likely that A pops up as a competitor during the later BX trials, leading the neural representation of A to distort. Studies that rely on the retrieval practice paradigm (or the reverse retrieval induced forgetting paradigms) to look at representational change [57] may also be useful, since one item in each pair (Rp+) undergoes retrieval practice whereas the other (Rp-) item does not (it is either restudied, or not presented at all during the retrieval practice phase). This arrangement makes it more likely that the Rp- item will pop up as a competitor in response to the Rp+ item than vice-versa.

Another possibility is to use neural activity to predict which item would distort. Since the representational change is rapid following competition, tracking the activity level of the competitor trial-by-trial could be used as a predictive measure for the distorting item.

The asymmetry of differentiation also has implications for how to measure differentiation effects more generally (even when you are not trying to detect asymmetry). In paradigms like the one used by Chanales et al. (2021) [102], our model predicts that only one item in each pair will repel. This implies that, if you measure repulsion by looking at *all* items (i.e., both items in each pair), you will be averaging across the pairmate that moved and the pairmate that stayed put—

including the latter group of items should weaken the measured effect. Statistically, it is more powerful to use designs where there is a way of predicting *a priori* which item will shift and which will stay, so you can focus on the subset of items that shift without "diluting" the analysis by including items that stay.

Limitations and open questions

Our model can account for differentiation and integration across several scenarios. We think this provides important computational support for the NMPH explanation of representational change, and useful predictions for future work. Nonetheless, there are several ways our model can be extended in the future.

We intentionally kept our model very simple in order to create a sandbox to explore the dynamics at play during learning. We used no initial study task for our model, and we instead opted for pre-wired connections in order to precisely label units as belonging to one pairmate or another. This simplification made it straightforward to explore the consequences of representational change—for instance, the asymmetric distortion effect was plainly visible, given that the hidden layer units for pairmates A and B were consistently assigned.

Rather than skipping initial stimulus learning and focusing on the first moment that competition happens, future work can investigate the trials leading up to this moment in the course of learning. For instance, our model of Schlichting et al. (2015) [11] instantiates interleaved vs blocked learning by changing the initial weights to the output X unit. Future models could instead show stimuli in a blocked or interleaved fashion, which should automatically lead to the kind of weights we pre-wired into the model.

This model could be expanded to be more biologically realistic (e.g., [25,29]). We have discussed how different inhibition levels in different brain areas could lead to varying results (**Fig 3.9D**), but building one model incorporating both neocortical and hippocampal regions would allow exploration of how these regions work together to support plasticity in the brain. Our simulations show that fast-learning regions like the hippocampus can support differentiation, whereas slower-learning regions like neocortex cannot support differentiation on their own. Thus, to the extent that differentiation effects have been observed in neocortex in the paradigms discussed here, we hypothesize that these effects are being propped up by top-down feedback from hippocampus; this could be tested (in a model and in experiments) by stimulating the hippocampus to disrupt activity, and measuring whether this eliminates cortical differentiation. Importantly, we also hypothesize that—in the longer term—hippocampus may be able to act as a teacher to support long-lasting differentiation within neocortex. While our simulations suggest that neocortex *on its own* cannot enact lasting differentiation via NMPH learning (due to neocortex's slow learning rate, which causes it to "relapse" and reabsorb formerly-shared units), differentiated hippocampal representations—once they are formed—may be able to act as a "pry bar" that operates on cortical representations to gradually push them apart. Another benefit of studying NMPH learning using a more biologically-realistic model is that we could simulate the contributions of different hippocampal subfields to representational change (e.g., contrasting the monosynaptic and trisynaptic pathways [25]). We will explore these ideas in future simulations.

Also, as noted earlier, our keep-things-simple approach led us to focus on NMPH learning in these simulations—in future work, we plan to explore how NMPH interacts with error-driven (supervised) learning. While supervised learning is not sufficient to explain effects modeled here, there is ample evidence that error-driven learning happens in the brain, and it may

interact with and/or complement the NMPH principles explored here in interesting ways. For example, in cases where NMPH fails to support differentiation of competing representations on its own, it may be possible to achieve differentiation of these representations via incremental error-driven learning (i.e., if error-driven learning is operating in a particular brain system, and it is not "drowned out" by other forms of learning like NMPH, then—with enough exposure—training a person to link similar stimuli to distinct outcomes should eventually result in the representations of those outcomes moving apart). We will also explore the influence of other biologically-inspired learning principles, including the principle of metaplasticity (whereby the transition points between strengthening and weakening are adjusted as a function of activity), which has previously been shown to play an important role in NMPH-style learning [53].

Lastly, several studies indicate that such representational change may be useful for reducing interference errors [102,10]. For instance, Favila et al. (2016) [10] showed that lower representational overlap in the hippocampus benefited subsequent learning on a different task. Such additional learning was outside the scope of our model, but future models could investigate this by adding more external inputs for additional learned associates.

Summary

The model presented in this paper provides a much-needed computational framework to support the NMPH account of representational change. By exploring different ways competitor activation can be modulated (by varying stimulus similarity, output similarity, learning conditions, or inhibitory dynamics), our model shows how subtle changes can affect whether differentiation or integration occurs. While there is more work to be done to explore the

consequences of this U-shaped learning function on representational change, we hope our model can be instrumental in framing future modeling and empirical research on learning.

Methods

Model architecture

We built the model in this research using the Leabra architecture in the Emergent neural network simulation software [108], written in Go. We built the general model as described in "Basic Network Properties," and adapted the parameters to model three individual studies. Some of the changes we made were structural changes to account for the different paradigms, as outlined in the above sections for each experiment. These structural changes had consequences on network dynamics that required various other parameters to be adjusted. This section will outline the details of these adjustments.

Activation and inhibitory dynamics

Emergent simplifies the discrete on-off firing of individual neurons into a rate code approximation where a unit's activity can range from 0 to 1, which reflects the overall spiking activity given the unit's current level of excitatory, inhibitory, and external inputs. A given unit M's excitatory input is a function of the activity of all other units N connected to it, weighted by the connection strength between M and N. This function is S-shaped from 0 to 1, where the steepness of that function is modulated by a *Gain* factor, which just multiplies the whole function. Lowering the *Gain* creates a more graded response, which enables moderate pop up of competitor units. The *Gain* for each layer is written in **Table 3.1** (*XX1 Gain)*, along with the

other activation and inhibitory parameters for these models. When the external input is clamped to a layer, a *Clamp Gain* factor is multiplied by the external input, modulating how strongly it should contribute to activity compared to the excitatory input.

A unit's activity is also modulated by the inhibitory dynamics in the layer. We implemented inhibition using an adaptation of the *k-winners take all* (kWTA) algorithm [39]. This algorithm imposes a "set point"-like behavior on inhibition, ensuring that only *k* units are allowed to activate in a layer. Typically in the kWTA algorithm, the units in a layer are "ranked" in terms of the amount of inhibition that would be needed to put that unit at the threshold of activating, given its current excitatory input. Then, inhibition is set such that only the *k* highest ranked units activate.

However, we found that we had to adapt the classic form of the kWTA algorithm. In several situations in our model, we wanted to allow more than *k* units to activate. It is useful for integration, for instance, to allow more than *k* units to activate in order to incorporate units belonging to both pairmates. In the conditions that lead to differentiation, allowing more than *k* units also helps circumvent a tie-breaker issue when adding new units to the damaged

Model	Layer	K	K Max	kWTA Point	Target Diff	Osc	XX1 Gain	Clamp Gain
Chanales	Output	6	15	0.95	0.05	0.115	30	—
Chanales	Hidden	6	10	0.75	0.03	0.11	100	—
Chanales	Category	1	∞	0.75	0.	0.	100	2
Chanales	Item	1	∞	0.95	0.2	0.22	100	0.3
Favila	Output	1	∞	0.75	0.03	0.07	100	—
Favila	Hidden	6	10	0.8	0.02	0.067	100	—
Favila	Category	1	∞	0.75	0	0	100	2
Favila	Item	1	∞	0.95	0.2	0.2	100	0.3
Schlichting	Output	1	∞	0.75	0.03	0.03	100	—
Schlichting	Hidden	6	10	0.8	0.02	0.0623	100	—
Schlichting	Category	1	∞	0.75	0	0	100	2
Schlichting	Item	1	∞	0.95	0.2	0.156	100	0.3

Table 3.1: Parameters for layer inhibitory and activation dynamics. *kWTA Point = a value between 0 and 1, which indicates how far toward the K+1th unit to place the current inhibitory level (the higher K Point, the lower the inhibition value). Osc = the amplitude of the oscillation function which multiplies the overall inhibition level of the layer. Note that for the model of Schlichting et al (2015)., we tested three different hidden layer oscillation amounts: 0.0623, .0525 and .09. XX1 Gain = the multiplier on the S-shaped activation function, where lower values means that activation will be more graded. Clamp Gain = multiplies the external input, modifying how strongly it contributes to the activation of the layer.*

competitor's hidden layer representation; if only *k* units are allowed to activate and the inhibition needed to put the k^{th} unit at threshold is very close to the value needed for the $k+1^{th}$ unit, the activation of the k^{th} unit would be very low. Consequently, we adjusted the kWTA algorithm to allow units that are tied in the ranking (within a threshold amount reflected by the *Target Diff* parameter) to activate as well. We also included a cap on the total number of units allowed to activate even with the tie (reflected by the *K Max* parameter). So if $k = 6$ and the *K Max* = 10, the top *k* units will be allowed to activate, as well as optionally any units more than 6 that are tied (within some threshold specified by *Target Diff*), but no more than 10.

Lastly, inhibition is modulated by oscillations, which are important to control the amount of competitor activity. The layer's inhibition is initially constant (determined by the above kWTA calculation), and then at cycle 125 is treated as a sine curve with a parameterized amplitude, until the end of the trial at cycle 200. The raising of inhibition does not have much impact on the network activity, but lowering the inhibition allows the competitors to activate. This allows competitor units to pop up when inhibition is lowered half way through the period of the sine wave, at cycle 163.

Projections between layers

The weight of any given connection between two units could range from 0 to 1. The two input layers are connected to the hidden layer, which is in turn connected to the output layer. The hidden layer also has recurrent connections. Generally the output layer does not have recurrent connections, except for our model of Chanales et al. (2021) [102] where we included recurrent output-to-output connections to simulate color selectivity. All projections were bidirectional—activity was able to flow forwards and backwards through the network. Parameters for each

bidirectional projection were identical except for the *Wt Scale*, which indicated the overall strength of the projection from one layer to another.

All layers that were connected were connected fully—that is, all units in one layer were connected to all units in the other layer. Most of these connections were random and low level, except for a set of pre-wired connections which were stronger, as described in "Basic Network Properties." Any pre-wired connections that differed between versions of the model are included in the sections of this paper on the set up of each model. All pre-wired, non-random connections were set to have a strength of .99 unless otherwise indicated. The parameters for these random connections are shown in **Table 3.2**.

The random low level connections represents one of two sources of variance— the other being the trial order for most simulations (described below in "Stimulus Presentation").

Learning

We used only unsupervised NMPH learning in this model since we wanted to see whether the NMPH was sufficient to produce the representational change seen in these experiments. Each of the projections in the model included learning, except for the added projection from output-to-output in the model of Chanales et al. (2021) [102].

Model	Projection	Wt Range	Wt Scale (Forwards / Backwards)
Chanales	Hidden ↔ Hidden	0.45 - 0.55	1.8 / 1.8
Chanales	Hidden ↔ Output	0.01 - 0.03	3.0 / 2.0
Chanales	Category ↔ Hidden	0.01 - 0.03	0.2 / 0.2
Chanales	Item ↔ Hidden	0.45 - 0.55	0.2 / 0.2
Chanales	Output ↔ Output	0.01 - 0.03	1.0 / 1.0
Favila	Hidden ↔ Hidden	0.45 - 0.55	1.8 / 1.8
Favila	Hidden ↔ Output	0.01 - 0.03	1.2 / 1.7
Favila	Category ↔ Hidden	0.01 - 0.03	0.1 / 0.1
Favila	Item ↔ Hidden	0.45 - 0.55	0.3 / 0.2
Schlichting	Hidden ↔ Hidden	0.45 - 0.55	1.9 / 1.9
Schlichting	Hidden ↔ Output	0.01 - 0.03	1.2 / 1.7
Schlichting	Category ↔ Hidden	0.01 - 0.03	0.2 / 0.1
Schlichting	Item ↔ Hidden	0.45 - 0.55	0.3 / 0.2

***Table 3.2: Projection parameters**. Wt Range = range of the uniform distribution used to initialize the random weights between each projection (range does not include the higher, pre-wired connections described in the text, which is set to .99 unless stated otherwise). Wt Scale = the scaling of the projection, in terms of an absolute multiplier on the weights in the projection.*

After each training trial, connection weights were adjusted based on the activity of the units at that point. For each pair of connected units, the coactivity is computed as the product of both units' short-term average activation. The U-shaped function is defined by five parameters (**Fig 3.10**), and is applied to this coactivity value to calculate the magnitude and direction of the weight change.

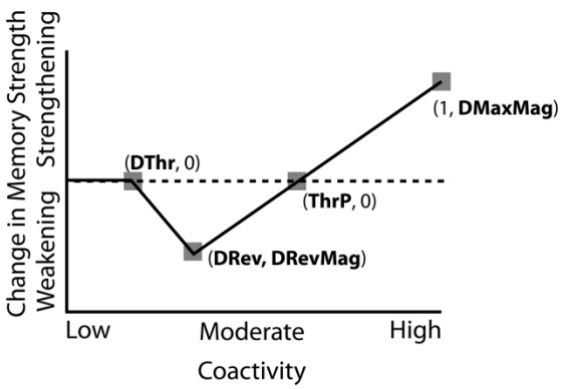


Figure 3.10: U-shaped learning function. *DThr, DRev, and ThrP are X axis coordinates. DRevMag and DMaxMag are Y axis coordinates, and indicate the amount of peak weakening or strengthening.*

These five parameters were separately set for each projection in each model so that when moderate competitor pop up occurred, the competitor-to-shared connections would sever. All reverse connections had identical U-shaped functions. Once the U-shaped function was set for a connection, those were the parameters used for all runs of the model, in all conditions. The parameters for each learning function can be found in **Table 3.3**.

Additionally, there is a parameter for the learning rate (LRate), which multiplies the change in weight output from the above function to calculate the final change in weight. In effect, then, a lower LRate makes the U of the U-shaped function more shallow and a higher LRate makes the U deeper. For all of our simulations LRate was set to 1 so the learning was not affected, except for the simulation where we explicitly adjusted the LRate parameter.

Stimulus presentation

Each model run included only two stimuli—pairmates A and B. We ran one pair per condition, and even though our input layers could be used for multiple conditions, we ran each

Model	Projection	DThr	DRev	DRevMag	ThrP	DMaxMag
Chanales	Hidden ↔ Hidden	0.15	0.24	-4.5	0.4	0.1
Chanales	Hidden ↔ Output	0.1	0.44	-10	0.6	1.5
Chanales	Category ↔ Hidden	0.2	0.3	-0.1	0.46	0.06
Chanales	Item ↔ Hidden	0.2	0.3	-2.5	0.46	0.3
Chanales	Output ↔ Output	0.53	0.6	-0.3	0.68	0.3
Favila	Hidden ↔ Hidden	0.11	0.23	-1.5	0.4	0.1
Favila	Hidden ↔ Output	0.11	0.23	-0.01	0.4	0.5
Favila	Category ↔ Hidden	0.2	0.3	-0.1	0.46	0.06
Favila	Item ↔ Hidden	0.215	0.4	-2.5	0.6	0.3
Schlichting	Hidden ↔ Hidden	0.11	0.23	-1.5	0.4	1
Schlichting	Hidden ↔ Output	0.11	0.23	-0.01	0.4	0.5
Schlichting	Category ↔ Hidden	0.2	0.3	-0.1	0.46	0.06
Schlichting	Item ↔ Hidden	0.11	0.23	-1.5	0.4	1

Table 3.3: Learning parameters*. All parameters are defined in Figure 10. All bidirectional connections used the same parameters for the U-shaped learning function.*

condition in the model separately for simplicity. We used the middle unit in the category layer and the middle column of two units in the item layer for pairmates A and B. The external inputs were thus the same for all models. Every epoch consisted of a single presentation of A and B, in a random order, except in our model for Schlichting et al. (2015) [11] where B was always shown first (and in the blocked condition, no A trials followed).

A test epoch was run before the first training epoch to attain a "baseline" measure of the representations of A and B, and then once more after each subsequent training epoch. The only difference between test and training epochs was that 1) test epochs included no oscillations, so we could get a pure guess for the state of the representation, and 2) connection weights were not adjusted at the end of test epochs. We ran the model for 20 train/test epochs, and the final pattern of unit activity for each test epoch was recorded and used for the analyses.

Practical advice for getting the model to show differentiation

As discussed above, differentiation is highly sensitive to activation dynamics. A couple of key events must occur on the second trial for differentiation to happen:

1. In trial 2, pairmate 1 units cannot be reactivated.
2. In trial 2, the hidden layer representation for pairmate 2 must still be accessible.

This seems straightforward, but meeting these conditions can be difficult. Sometimes the necessary parameter adjustment to satisfy #1 make #2 harder to satisfy, and vice versa. Most of the failures to differentiate that we observed happened if these two conditions were not fully met. (This indicates an important caveat to our model—if some of the parameters are slightly different, it is possible to go straight from "no change" to "integration" as competitor activity is increased.)

In trial 2, pairmate 1 units cannot be reactivated

Even if some weakening of competitor-to-shared units has occurred on trial 1, it is possible for pairmate 1 units to activate on trial 2. If this happens, any initial differentiation effects are undone. This happens if the excitation for pairmate 1 units remains too high, and there are several potential causes.

We have discussed one potential cause here, when learning rate is low and the competitor-to-shared connections are only mildly severed on trial 1. Additionally, if only some (but not all) competitor-to-shared connections weaken, the intact connections may reactivate pairmate 1 on a trial 2, undoing any initial differentiation. There are several situations in which we see this. If variation in the random connection weights is too high, not all competitor units will pop up the same amount. It is possible, then, to get a situation where not all competitor-to-competitor coactivities are less than all competitor-to-shared coactivities, meaning it is impossible to preserve within-competitor links while also severing links between competitor-to-

shared units (**Fig 3.11**). This non-uniform severing can also occur across different projections; if the competitor pop up is different across layers, it is possible that some projections will sever all competitor-to-shared connections while other projections will not, leading pairmate 1 to continue to receive high excitation on trial 2.

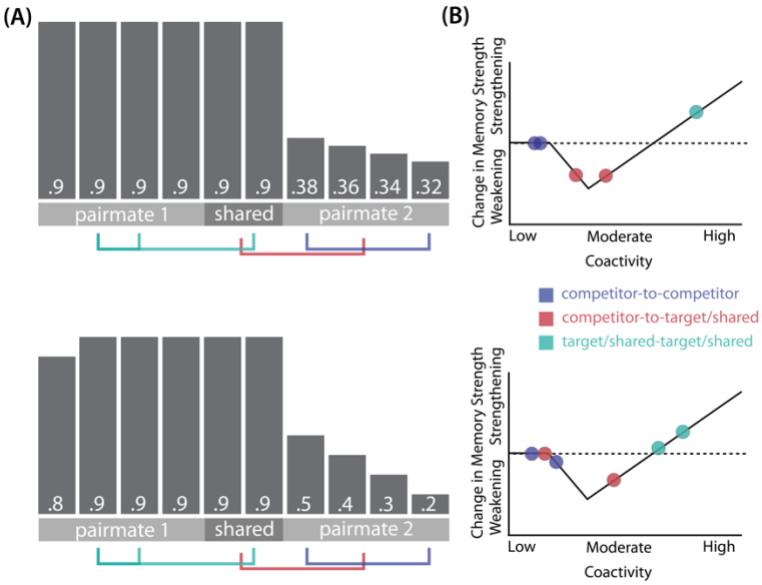


Figure 3.11: How subtle changes to activity can affect representational change. (A) *Two sample activity patterns in the hidden layer during the first trial are shown. Units are labelled as belonging to either pairmate 1, pairmate 2, or both, and vertical bars indicate activation level (along with activation values). A moderate amount of pop up of pairmate 2 occurs, which should lead to differentiation if the competitor-to-shared units appropriately sever. While the two patterns are very similar, the bottom pattern has slightly more variable activity, which could arise if the level of random noise in the hidden-to-hidden connections is higher.* ***(B)*** *A U-shaped function for each activity pattern is shown, with the coactivity for hidden-to-hidden connections plotted along the X axis. As shown in the top example, a U-shaped function exists such that all competitor-to-competitor coactivities are lower than all competitor-to-target/shared coactivities, which are in turn lower than all target/shared-to-target/shared coactivities. This means that it is possible to preserve the competitor-to-competitor connections while severing connections to the shared units, leading to differentiation. However, in the bottom example, the added noise means that the ordering of these coactivities is not consistent, and it is not possible to fully preserve all competitor-to-competitor connections while severing all competitor-to-target/shared connections. Consequently, even though the competitor units moderately pop up, this pattern of activity may or may not lead to differentiation.*

There are some connections that will always send excitation to pairmate 1 on trial 2, so some amount of excitation of the hidden pairmate 1 units is unavoidable. One projection for which this is true for all models is the category-to-hidden connection, since the category unit sends excitation to both A and B throughout all trials. Some of the models also have other units that maintain a connection to both A and B—the hidden-to-output connections in the models of Favila et al. (2016) [10] and Schlichting et al. (2015) [11], for instance. Lowering the initial connection strength between these layers can help prevent this, so that pairmate 1 receives less excitation.

Lastly, one simple fix that may work is just to change the oscillation amount, which can shift the placement of competitor-to-shared coactivity on the U-shaped function. If the oscillation is not set to the appropriate amount, the competitor units may activate to a level that does not cause appropriate severing of competitor-to-share connections, leading the shared units to activate on trial 2. We can see in the results of our model of Schlichting et al. (2015) [11]—when the oscillation amount alone is adjusted, this can change the outcome of representational change (**Fig 3.9A** and **3.9D**).

In trial 2, the hidden layer representation for pairmate 2 must still be accessible

If the competitor-to-competitor units are severed too much, this means that—when it is time to return to a trial where this former-competitor is the target—its hidden representation will not be accessible. When this happens, the remaining connection from the category layer will take over, and A units will activate. This tends to lead to the asymmetric form of integration, where the damaged pairmate will instead connect to the units assigned to its competitor.

Any parameter adjustment that lowers the amount of pop up is helpful, since that moves the competitor-to-competitor coactivities to the left on the U-shaped function. For instance, decreasing the oscillation amount may be needed to make sure competitor-to-competitor units are spared in trial 1. However, this can be difficult, since if the oscillation amount is decreased it may also cause competitor-to-shared connections to be spared as well, we discussed in #1.

Conclusion

Learning involves changing the brain as a result of experience in order to help us interact with the world better in the future. Neural representations of some memories move together (integration), allowing us to generalize, whereas other memories move apart (differentiation), allowing us to discriminate. In this way, our memory system performs a delicate balancing act to create the complicated and vast array of knowledge we hold. But ultimately, learning itself is a very simple process: strengthening or weakening individual neural connections. How can a simple learning rule on the level of individual neural connections account for both differentiation and integration of entire memories, and what factors lead to both outcomes?

We first presented a unifying theory for this array of findings: The NMPH, which says that moderately active memories are weakened and strongly active memories are strengthened, can scale up in such a way that moderately active competing memories are differentiated and strongly active competing memories are integrated. We then explored this through behavioral experiments, which were inconclusive in and of themselves but provide insights that will help guide research into this area. Lastly, we built a neural network model that instantiates this U-shaped, unsupervised learning rule, to provide computational evidence that the NMPH can explain results whereby similar memories lead to differentiation. This both validated the theory we presented in Chapter 1, and provided several novel, testable predictions—such that differentiation is rapid and asymmetric.

The simulation results presented in Chapter 3 also shed some light on the work presented in Chapter 2, and why we did not observe distortion effects in these studies. The results of the behavioral studies we ran suffered from high swap rates, which may be an example of the kind of asymmetric integration we found in the model. Our model also showed that differentiation is

highly sensitive to many parameters in the model, and if every parameter is not precisely tuned, differentiation may fail and instead lead to asymmetric integration. It is possible this is what happened in our studies.

Moreover, our simulation shows it is critical not to reactivate pairmate 1 after connections have been severed, otherwise differentiation may be undone. This may additionally explain some of the swaps in our behavioral studies. In our simulations, we discussed that it was important to keep the influence of shared features to a minimum in order for differentiation to occur. This may mean the idea we presented in Chapter 2 of cueing-from-the-side may be an important path to pursue for future research.

Additionally, in light of the predicted asymmetry of differentiation, the retrieval practice design we used (where competitive retrieval only happens for the Rp+ item) may be an ideal since it leads to clear predictions about which items should differentiate. Without this, it may be hard, in practice, to identify which item within a pair should differentiate, since many factors may impact when the critical moment of sufficient competition occurs to trigger differentiation.

There are several outstanding questions that future work can explore, beyond the specific modeling questions discussed above. For instance, what are the most sensitive ways to measure the behavioral consequences of neural differentiation and integration? This dissertation discusses the paradigm of Chanales et al. (2021) [102] in depth, where color was used to measure attraction and repulsion effects, but further work could explore other ways these distortion effects may be observed.

Future work can also explore the role offline learning (e.g., during sleep or quiet wake) plays in representational change, and how the NMPH contributes. Neural activity is reactivated (i.e., replayed) during sleep [99]. Unsupervised learning may be especially important (relative to

supervised learning) during sleep, since the brain is cut off from the external world and sensory input cannot provide a "target" for learning.

There may additionally be a potential for the NMPH to be leveraged to develop new clinical treatments. Several clinical conditions are characterized by insufficient neural differentiation (e.g., of phonological representations, in the case of developmental dyslexics; and visual representations, in the case of patients with temporal lobe damage). A deeper understanding of representational change may provide new and better ways of re-differentiating collapsed representations and ameliorating associated cognitive deficits. There is evidence showing that differentiating phonological representations can help remediate dyslexia [112]. A deeper understanding of these learning dynamics may provide new and better ways of re-differentiating these collapsed representations and ameliorating the associated cognitive deficits.

Future work may find neurofeedback to be especially useful in order to consistently evoke the levels of moderate activation needed for differentiation. Real-time methods could be harnessed to manipulate whether integration or differentiation occurs. For instance, the multiple object tracking procedure has been shown to be a parametrically adjustable way of interfering with memory activation [74]. Neurofeedback could be used to adjust the specific amount of activation needed to move memory activation into the dip of the U-shaped function.

This work helps advance the understanding of how we learn, bridging the gap between an array of findings and provides a novel framework for representational change. The questions that stem from this work will hopefully present targets for future work in the field.

References

1 Norman, K.A. *et al.* (2006) How inhibitory oscillations can train neural networks and punish competitors. *Neural Comput.* 18, 1577–1610

2 McClelland, J.L. *et al.* (1995) Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* 102, 419–457

3 Lewis-Peacock, J.A. and Norman, K.A. (2013) Multi-voxel pattern analysis of fMRI data. *Cogn. Neurosci. V* 20, 304–313

4 Kriegeskorte, N. *et al.* (2008) Representational similarity analysis–connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2, 4

5 Rumelhart, D.E. *et al.* (1986) Learning internal representations by error propogation. In *Parallel Distributed Processing: Explorations in the Microstructure of Vognition, Vol. 1: Foundations* pp. 318–362, MIT Press

6 Gluck, M.A. and Myers, C.E. (1993) Hippocampal mediation of stimulus representation–A computational theory. *Hippocampus* 3, 491–516

7 Schapiro, A.C. *et al.* (2013) Neural representations of events arise from temporal community structure. *Nat. Neurosci.* 16, 486–492

8 Schapiro, A.C. *et al.* (2016) Statistical learning of temporal community structure in the hippocampus. *Hippocampus* 26, 3–8

9 Tompary, A. and Davachi, L. (2017) Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron* 96, 228–241.e5

10 Favila, S.E. *et al.* (2016) Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nat. Commun.* 7, 11066

11 Schlichting, M.L. *et al.* (2015) Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nat. Commun.* 6, 8151

12 Dimsdale-Zucker, H.R. *et al.* (2018) CA1 and CA3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields. *Nat. Commun.* 9,

13 Lecun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444

14 Yamins, D.L.K. and DiCarlo, J.J. (2016) Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365

15 Khaligh-Razavi, S.M. and Kriegeskorte, N. (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* 10,

16 Scellier, B. and Bengio, Y. (2017) Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Front. Comput. Neurosci.* 11,

17 Reilly, R.C.O. (1996) Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Comput.* 8, 895–938

18 O'Reilly, R.C. *et al.* (2017) Deep predictive learning: a comprehensive model of three visual streams. *arXiv* at <http://arxiv.org/abs/1709.04654>

19 Keller, G.B. and Mrsic-Flogel, T.D. (2018) Perspective predictive processing: A canonical cortical computation. *Neuron* 100, 424–435

20 Whittington, J.C.R. and Bogacz, R. (2019) Theories of error back-propagation in the brain. *Trends Cogn. Sci.* 23, 235–250

21 Richards, B.A. and Lillicrap, T.P. (2019) Dendritic solutions to the credit assignment problem. *Curr. Opin. Neurobiol.* 54, 28–36
22 McClelland, J.L. and Rogers, T.T. (2003) The parallel distributed processing approach to semantic cognition. *Nat. Rev. Neurosci.* 4, 310–322
23 Milivojevic, B. *et al.* (2015) Insight reconfigures hippocampal-prefrontal memories. *Curr. Biol.* 25, 821–830
24 Collin, S.H.P. *et al.* (2015) Memory hierarchies map onto the hippocampal long axis in humans. *Nat. Neurosci.* 18, 1562–1564
25 Schapiro, A.C. *et al.* (2017) Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philos. Trans. R. Soc. B Biol. Sci.* 372, 20160049
26 Ballard, I.C. *et al.* (2019) Hippocampal pattern separation supports reinforcement learning. *Nat. Commun.* 10, 1073
27 Marr, D. (1971) Simple memory: A theory for archicortex. *Philos. Trans. R. Soc. B Biol. Sci.* 262, 23–81
28 O'Reilly, R.C. and McClelland, J.L. (1994) Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus* 4, 661–682
29 Norman, K.A. and O'Reilly, R.C. (2003) Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychol. Rev.* 110, 611–646
30 Yassa, M.A. and Stark, C.E.L. (2011) Pattern separation in the hippocampus. *Trends Neurosci.* 34, 515–525
31 Chanales, A.J.H. *et al.* (2017) Overlap among spatial memories triggers repulsion of hippocampal representations. *Curr. Biol.* 27, 2307–2317
32 Duncan, K.D. and Schlichting, M.L. (2018) Hippocampal representations as a function of time, subregion, and brain state. *Neurobiol. Learn. Mem.* 153, 40–56
33 Rafidi, N.S. *et al.* (2018) Reductions in retrieval competition predict the benefit of repeated testing. *Sci. Rep.* 8, 1–12
34 Antony, J.W. *et al.* (2017) Retrieval as a fast route to memory consolidation. *Trends Cogn. Sci.* 21, 573–576
35 Rowland, C.. (2014) The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychol. Bull.* 140,
36 van den Broek, G. *et al.* (2016) Neurocognitive mechanisms of the "testing effect": A review. *Trends Neurosci. Educ.* 5, 52–66
37 Karpicke, J.D. and Roediger, H.L. (2008) The critical importance of retrieval for learning. *Science (80-.).* 319, 966–969
38 Grossberg, S. (1976) Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biol. Cybern.* 23,
39 O'Reilly, R.C. and Munakata, Y. (2000) *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*, The MIT Pr.
40 Buonomano, D. V. and Merzenich, M.M. (1998) Cortical plasticity: From synapses to maps. *Annu. Rev. Neurosci.* 21, 149–186
41 McClelland, J.L. *et al.* (2002) Teaching the /r/-/l/ discrimination to Japanese adults: Behavioral and neural aspects. *Physiol. Behav.* 77, 657–662
42 Hall, G. (1996) Learning about associatively activated stimulus representations: Implications for acquired equivalence and perceptual learning. *Anim. Learn. Behav.* 24,

233–255
43 McCandliss, B.D. *et al.* (2002) Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cogn. Affect. Behav. Neurosci.* 2, 89–108
44 Symonds, M. and Hall, G. (1995) Perceptual learning in flavor aversion conditioning: Roles of stimulus comparison and latent inhibition of common stimulus elements. *Learn. Motiv.* 26, 203–219
45 Clark, S.A. *et al.* (1988) Receptive fields in the body-surface map in adult cortex defined by temporally correlated inputs. *Nature* 332, 444–445
46 Bienenstock, E.L. *et al.* (1982) Theory for the development of neuron selectivity: Orientation Sspecificity and binocular interaction in visual cortex. *J. Neurosci.* 2, 32–48
47 Cooper, L.N. (2004) *Theory of cortical plasticity*, World Scientific.
48 Diederich, S. and Opper, M. (1987) Learning of correlated patterns in spin-glass networks by local learning rules. *Phys. Rev. Lett.* 58, 949–952
49 Detre, G.J. *et al.* (2013) Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia* 51, 2371–2388
50 Newman, E.L. and Norman, K.A. (2010) Moderate excitation leads to weakening of perceptual representations. *Cereb. Cortex* 20, 2760–2770
51 Artola, A. *et al.* (1990) Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature* 347, 69–72
52 Hansel, C. *et al.* (1996) Different threshold levels of postsynaptic [Ca2+]i have to be reached to induce LTP and LTD in neocortical pyramidal cells. *J. Physiol. - Paris* 90, 317–319
53 Bear, M.F. (2003) Bidirectional synaptic plasticity: From theory to reality. *Philos. Trans. R. Soc. B Biol. Sci.* 358, 649–655
54 Lewis-Peacock, J.A. and Norman, K.A. (2014) Competition between items in working memory leads to forgetting. *Nat. Commun.* 5, 5768
55 Kim, G. *et al.* (2014) Pruning of memories by context-based prediction error. *Proc. Natl. Acad. Sci.* 111, 8997–9002
56 Wang, X.T.H. *et al.* (2019) More is less: Increased processing of unwanted memories facilitates forgetting. *J Neurosci* 39, 3551–3560
57 Hulbert, J.C. and Norman, K.A. (2015) Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice. *Cereb. Cortex* 25, 3994–4008
58 Paller, K.A. *et al.* (2019) Replay-based consolidation governs enduring memory storage. *Cogn. Neurosci. VI MIT Press*
59 Norman, K.A. *et al.* (2007) A neural network model of retrieval-induced forgetting. *Psychol. Rev.* 114, 887–953
60 Anderson, M.C. *et al.* (1994) Remembering can cause forgetting: Retrieval dynamics in long-term memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 20, 1063–1087
61 Wimber, M. *et al.* (2015) Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nat. Neurosci.* 18, 582–589
62 Murayama, K. *et al.* (2014) Forgetting as a consequence of retrieval: A meta-analytic review of retrieval-induced forgetting. DOI: 10.1037/a0037505
63 Storm, B.C. *et al.* (2008) Accelerated relearning after retrieval-induced forgetting: The benefit of being forgotten. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 230–236

64 Chan, J.C.K. (2009) When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *J. Mem. Lang.* 61, 153–170
65 Zeithamova, D. *et al.* (2012) Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron* 75, 168–179
66 Kim, G. *et al.* (2017) Neural differentiation of incorrectly predicted memories. *J. Neurosci.* 37, 2022–2031
67 Greve, A. (2018) Opinion: Neural differentiation of incorrectly predicted memories. *Front. Hum. Neurosci.* 12,
68 Schapiro, A.C. *et al.* (2012) Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Curr. Biol.* 22, 1622–1627
69 Barnes, C.A. *et al.* (1990) Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Prog. Brain Res.* 83, 287–300
70 Keinath, A.T. *et al.* (2014) Precise spatial coding is preserved along the longitudinal hippocampal axis. *Hippocampus* 24, 1533–1548
71 Myung, J.I. *et al.* (2013) A tutorial on adaptive design optimization. *J Math Psychol.* 57, 53–67
72 Sitaram, R. *et al.* (2017) Closed-loop brain training; the science of neurofeedback. *Nat. Rev. Neurosci.* 18, 86–100
73 DeBettencourt, M.T. *et al.* (2015) Closed-loop training of attention with real-time brain imaging. *Nat. Neurosci.* 18, 470–475
74 Poppenk, J. and Norman, K.A. (2017) Multiple-object tracking as a tool for parametrically modulating memory reactivation. *J. Cogn. Neurosci.* 29, 1339–1354
75 deBettencourt, M.T. *et al.* (2019) Neurofeedback helps to reveal a relationship between context reinstatement and memory retrieval. *Neuroimage*
76 Tang, H. *et al.* (2018) Recurrent computations for visual pattern completion. *Proc. Natl. Acad. Sci.* 115, 8835–8840
77 O'Reilly, R.C. *et al.* (2013) Recurrent processing during object recognition. *Front. Psychol.* 4, 1–14
78 O'Reilly, R. *et al.* (2015) The leabra cognitive architecture: how to play 20 principles with nature and win! *Oxford Handb. Cogn. Sci.* DOI: 10.1186/1472-6947-12-132
79 Hinton, G.E. and Salakhutdinov, R.R. (2006) Reducing the dimensionality of data with neural networks. *Science (80-.).* 313, 504–507
80 Love, B.C. *et al.* (2004) SUSTAIN: A network model of category learning. *Psychol. Rev.* 111, 309–332
81 Mack, M.L. *et al.* (2016) Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proc. Natl. Acad. Sci.* 113, 13203–13208
82 Klimesch, W. (1999) EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev.* 29, 169–195
83 Osipova, D. *et al.* (2006) Theta and gamma oscillations predict encoding and retrieval of declarative memory. *J. Neurosci.* 26, 7523–7531
84 Sederberg, P.B. *et al.* (2003) Theta and gamma oscillations during encoding predict subsequent recall. *J. Neurosci.* 23, 10809–14
85 Kuhl, B.A. *et al.* (2007) Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nat. Neurosci.* 10, 908–914
86 Newman, E.L. *et al.* (2012) Cholinergic modulation of cognitive processing: insights

drawn from computational models. *Front. Behav. Neurosci.* 6, 24
87 O'Reilly, R.C. *et al.* (2012) *Computational cognitive neuroscience (1st edition), wiki book*, 4th Editio.
88 Kirkwood, A. *et al.* Experience-dependent modification of synaptic plasticity in visual cortex. , *Nature*, 381. (1996) , 526–528
89 Anderson, M. and Green, C. (2001) Suppressing unwanted memories by executive control. *Nature* 410, 131–134
90 Komatsu, H. *et al.* (1992) Color selectivity of neurons in the inferior temporal cortex of the awake macaque monkey. *J. Neurosci.* 12, 408–24
91 Hanazawa, A. *et al.* (2000) Neural selectivity for hue and saturation of colour in the primary visual cortex of the monkey. *Eur. J. Neurosci.* 12, 1753–1763
92 Ciranni, M.A. and Shimamura, A.P. (1999) Retrieval-induced forgetting in episodic memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 25, 1403–1414
93 Zhang, W. and Luck, S.J. (2008) Discrete fixed-resolution representations in visual working memory. *Nature* 453, 233–235
94 Fan, J.E. and Turk-Browne, N.B. (2013) Internal attention to features in visual short-term memory guides object learning. *Cognition* 129, 292–308
95 Wilson, M.A. and McNaughton, B.L. (1994) Reactivation of hippocampal ensemble memories during sleep. *Science (80-.).* 29, 676–679
96 Ji, D. and Wilson, M.A. (2007) Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat. Neurosci.* 10, 100–107
97 Ellenbogen, J.M. *et al.* (2007) Human relational memory requires time and sleep.
98 Antony, J.W. *et al.* (2018) Competitive learning modulates memory consolidation during sleep. *Neurobiol. Learn. Mem.* 155, 216–230
99 Baran, B. *et al.* (2010) REM-dependent repair of competitive memory suppression. *Exp. Brain Res.* 203, 471–477
100 Norman, K.A. *et al.* (2005) Methods for reducing interference in the Complementary Learning Systems model: Oscillating inhibition and autonomous memory rehearsal. *Neural Networks* 18, 1212–1228
101 Singh, D. *et al.* (2022) A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation.
102 Chanales, A.J.H. *et al.* (2021) Adaptive repulsion of long-term memory representations is triggered by event similarity. *Psychol. Sci.* 32, 705–720
103 Zhao, Y. *et al.* (2021) Adaptive memory distortions are predicted by feature representations in parietal cortex. *J. Neurosci.*
104 Op De Beeck, H. *et al.* (2001) Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat. Neurosci.* 4, 1244–1252
105 Wammes, J.D. *et al.* (2019) Synthesizing images with deep neural networks to manipulate representational similarity and induce representational change. *J. Vis.* 19,
106 Li, A.Y. *et al.* (2019) The validated circular shape space: Quantifying the visual similarity of shape. *J. Exp. Psychol. Gen.* 149, 949
107 Ritvo, V.J.H. *et al.* (2019) Nonmonotonic plasticity: How memory retrieval drives learning. *Trends Cogn. Sci.* 23, 726–742
108 Aisa, B. *et al.* (2008) The Emergent neural modeling system. *Neural Networks* 21, 1146–1152
109 Wanjia, G. *et al.* Abrupt hippocampal remapping signals resolution of memory

interference. *Nat. Commun.* DOI: 10.1038/s41467-021-25126-0
110 Molitor, R.J. *et al.* (2021) Memory reactivation during learning simultaneously promotes Dentate Gyrus / CA23 pattern differentiation and CA1 memory integration. 41, 726–738
111 Wammes, J. *et al.* (2022) Increasing stimulus similarity drives nonmonotonic representational change in hippocampus. *Elife* 11, 1–23
112 Temple, E. *et al.* (2003) Neural deficits in children with dyslexia ameliorated by behavioral remediation: Evidence from functional MRI. *Proc. Natl. Acad. Sci.* 100, 2860–2865