MEASURING LONG-TERM MEMORIES AT THE FEATURE LEVEL REVEALS MECHANISMS OF INTERFERENCE RESOLUTION

by

MAXWELL L. DRASCHER

A DISSERTATION

Presented to the Department of Psychology
and the Division of Graduate Studies of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

March 2023

DISSERTATION APPROVAL PAGE

Student: Maxwell L. Drascher

Title: Measuring Long-term Memories at the Feature Level Reveals Mechanisms of Interference Resolution

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Psychology by:

Brice Kuhl	Chairperson
Ulrich Mayr	Core Member
Margaret Sereno	Core Member
James Murray	Institutional Representative

and

Krista Chronister	Vice Provost for Graduate Studies

Original approval signatures are on file with the University of Oregon Division of Graduate Studies.

Degree awarded March 2023

DISSERTATION ABSTRACT

Maxwell L. Drascher

Doctor of Philosophy

Department of Psychology

March 2023

Title: Measuring Long-term Memories at the Feature Level Reveals Mechanisms of Interference Resolution

When memories share similar features, this can lead to interference, and ultimately forgetting. At the same time, many highly similar memories are remembered vividly for years to come. Understanding what causes interference and how it is overcome is key to understanding the vast human memory capacity. One unresolved challenge is that interference has primarily been studied with dichotomous measures of memory ("remembered", "forgotten"). This limits our understanding because memories are not all-or-none, they are comprised of multiple features, each of which can be recalled with different levels of detail or bias. In order to investigate this issue, this dissertation focuses on the use of face stimuli. Faces are a unique class of stimuli for studying memory interference in that they are readily parameterizable and humans are experts at perceiving them. This means that they can be manipulated to be similar enough to cause interference, but subtle differences can also be stored and later probed from long-term memory. This dissertation develops a methodology to create synthetic faces that can be manipulated and probed along a set of perceptually-important feature dimensions. This development process included documenting face landmark positions, sorting faces based on perceived similarity, and collecting subjective ratings on a corpus of 1,148 face images. In a series of three experiments, I then applied this novel methodology to understand how memories change at the feature level when there is interference between highly similar memories. I found two memory changes that specifically occurred when there was interference between highly similar stimuli: (1) during recollection there was a bias to exaggerate the subtle differences and (2) distinguishing features were recalled with greater consistency. Critically, these memory changes were adaptive in that they were associated with less interference-related errors. Finally, in a separate fMRI experiment, I used the same corpus of faces and feature dimensions to reconstruct faces based on patterns of fMRI activity evoked while viewing them. I argue that this approach can be utilized in the future to measure neural representational changes during interference resolution. Together our findings provide important insights into how the memory system resolves interference between highly similar memories.

This dissertation includes previously published and unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Maxwell L. Drascher

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Skidmore College, Saratoga Springs

DEGREES AWARDED:

Doctor of Philosophy, Psychology, 2023, University of Oregon
Master of Science, Psychology, 2017, University of Oregon
Bachelor of Arts, Psychology, 2012, Skidmore College

AREAS OF SPECIAL INTEREST:

Cognitive Neuroscience

PROFESSIONAL EXPERIENCE:

Graduate Research & Teaching Assistant, University of Oregon,
September 2016 – March 2023

Project Manager, Claremont Graduate University,
August 2015 – August 2016

Research Specialist, Princeton University,
June 2014 – July 2015

Research Assistant, UMASS Boston,
August 2012 – January 2014

GRANTS, AWARDS, AND HONORS:

NSF GRFP Honorable Mention, University of Oregon, 2018

First Year Fellowship, University of Oregon, 2016

Psi Chi, Skidmore College, 2011

Periclean Honors Forum, Skidmore College, 2008

PUBLICATIONS:

Drascher, M. L., & Kuhl, B. A. (2022). Long-term memory interference is resolved via repulsion and precision along diagnostic memory dimensions. *Psychonomic Bulletin & Review*, 1-15.

Chanales, A. J., Tremblay-McGaw, A. G., **Drascher, M. L.**, & Kuhl, B. A. (2021). Adaptive repulsion of long-term memory representations is triggered by event similarity. *Psychological science*, *32*(5), 705-720.

Siperstein, G. N., Parker, R. C., & **Drascher, M.L.** (2013). National snapshot of adults with intellectual disabilities in the labor force. *Journal of Vocational Rehabilitation*, 39(3), 157-165.

ACKNOWLEDGMENTS

This dissertation would not be here without the many people in my life. I am very grateful for the support I have received. I will highlight certain people, but there are too many to name.

I would first like to thank my advisor, Dr. Brice Kuhl. Dr. Kuhl has provided me with countless hours of advice and wisdom throughout my time at the University of Oregon. Dr. Kuhl's scientific perspective is all over this document. I am very grateful for all the support he has given me.

I would also like to thank the other members of my committee: Dr. Ulrich Mayr, Dr. Margaret Sereno, and Dr. James Murray. Thank you for your time, attention, and valuable insight. Thanking my committee would not be complete without a special mention of a member who sadly passed away, Dr. Sarah Dubrow. Dr. Dubrow made my experience as a graduate student richer and more joyful. I hope that hints of her thinking can be seen in this dissertation.

I would also like to thank the members of the Kuhl lab, past and present. Several people directly played a role in this research. I would like to specifically thank Dr. Hongmi Lee, Dr. Nicole Long, and Sarah Sweigart for their roles early in this research. I would also like to thank Alex Tremblay-McGaw for her overall support and for her role in landmarking face stimuli and collecting much of the data. Thank you to Paul Keene for his role in data analysis. Many other members of the lab have helped me, guided my thinking, and provided support. Thank you to everyone else in the lab who I neglected to highlight here.

I would also like to thank the staff both in the Psychology Department and at the Lewis Center for Neuroimaging for providing facilities, equipment, training, and support integral to collection of the data included in this project. I greatly appreciate all of the assistance they have provided to me.

It is also important to thank one of the most influential people on my journey here, my

undergraduate thesis advisor, Dr. Hugh Foley. Dr. Foley gave me confidence and helped instill a love of research in me. He has always been very supportive of me. It is remarkable how much of this work parallels what I worked on with him over ten years ago.

Importantly, I would never have made it to this point without the support of my loved ones. Thank you for not giving up on me. Thank you to my parents. And thank you to my amazing partner, Kathy Padgett.

This research was supported in part by NSF CAREER Award (BCS-1752921) and NIH-NINDS R01 (NS107727) awarded to Dr. Kuhl.

For Mom.
Thank you for believing in me.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Chapter I

INTRODUCTION

Introduction: how do we remember highly similar information?

The capacity of the human memory system is seemingly limitless (Brady et al., 2008). Yet, we also quite often forget. One of the central problems in memory research is understanding the reasons and circumstances that distinguish a memory that will be successfully recalled and one that will be forgotten (Anderson, 2003; Anderson et al., 1994; Anderson & Spellman, 1995; Crowder, 2014; Fawcett & Hulbert, 2020; Smith & Hunt, 2000). There is a lot we already understand about challenges to memory, in particular how similarity between memories can lead to interference, which increases the chance of forgetting (Anderson & Neely, 1996). However, in part due to methodological barriers, most of the progress in this research has come from studies that focus only on whether a memory is remembered or forgotten (Cooper & Ritchey, 2019). In contrast, relatively little attention has been paid to subtler changes to memory—at the individual *feature* level. This is critical, because as I will argue, measuring changes to a memory's features may be key to understanding the impact of interference, how interference is resolved, and ultimately the vast human memory capacity.

This dissertation focuses on memories as a multi-dimensional constellation of individual features (Cooper & Ritchey, 2019; Horner & Burgess, 2013; Horner & Burgess, 2014; Xue, 2018). For example, when you meet someone new, you form a memory of that person that consists of specific pieces of information or features (e.g. eye color or the shirt they were wearing). A week later, given a cue (e.g. their name), you may be able to retrieve that memory.

Critically, each of the features you perceived when first meeting them is unlikely to be retrieved with the same accuracy or level of detail (with some forgotten).

The central challenge that this dissertation addresses is applying this perspective of memory to the study of episodic memory interference. In order to study interference, we need a class of stimuli that can be manipulated to be sufficiently similar to cause interface, while distinct enough to be retrieved from memory. The features of these stimuli also need to be measurable along continuous dimensions that allow for the ability to track subtle memory changes. This dissertation focuses on first developing a methodology that meets these criteria (Chapter 2). In human research, faces are uniquely suited for this purpose. I proceed to show two applications of this methodology—both in a behavioral paradigm (Chapter 3) and in a neuroimaging paradigm (Chapter 4). I conclude with a discussion of what has been learned as well as potential future applications (Chapter 5). To preview, I found two distinct changes at the level of individual features of memory, that each may play a role in resolving interference, ultimately making it possible for humans to remember so many highly similar pieces of information.

The rest of this introductory chapter summarizes important background information. In the next section, I highlight some of what is already known about memory interference (Memory interference background). I then proceed to discuss the idea of adaptive memory distortions, the notion that deviations from perfectly accurate memories can be advantageous for navigating the world (Adaptive memory changes as a route to interference resolution). That perspective helps inform why it is important to measure memories at the feature-level. Here, I also introduce a computational model (Hulbert & Norman, 2015) that motivated much of this work. This model proposes a theory of how memory features may change as an adaptive response to interference. The next section provides background on how the features of memory have previously been measured, both through behavioral probes and fMRI analysis (Measuring the feature space of memory). This dissertation represents an innovation in the behavioral measure of feature memory and presents a path towards innovation in the measure of feature memory based on fMRI activity. I conclude with a brief preview of the rest of the dissertation (Goal and structure of the dissertation).

Memory interference background

Cognitive perspectives on memory interference

Every day we encounter moments, items, or events that are quite similar to ones we've seen before. For example, it is a cliché to note that you can't remember what you ate for breakfast. This is because the memory of your breakfast from this morning is probably quite similar to other mornings. In contrast, you may have no trouble recalling what you ate at the restaurant you visited for the first time last week as there are no other memories of meals at the same restaurant. The disruption in our memory system caused by another similar, *competing* memory, such as yesterday's breakfast, is known as memory interference (Anderson & Neely, 1996; Anderson & Spellman, 1995; Smith & Hunt, 2000).

There are several theories about the mechanisms underlying memory interference. That is, how is the architecture of our memory system organized such that similar items would create disruptions. Some theoretical perspectives focus on retrieval, assuming that the competing items are stored in long-term memory. Under this perspective, memory interference reflects a disruption in the ability retrieve those memories (Rajsic et al., 2017; Tulving, 1974). For example, in cases where memories share a common cue, the memory with the strongest association with the cue will tend to win out; further, it may actually suppress the association between the cue and the competing item (Anderson & Neely, 1996; Anderson et al., 1994; Gillund & Shiffrin, 1984; Melton & Irwin, 1940; Rundus, 1973). Alternatively, the memories may remain retrievable, but competition can create binding errors where a cue from one item is errantly associated with a competitor. In this case, we would expect to not only see memory errors, but specifically "swap" errors where competitors are preferentially recalled (Bays et al., 2009). Relatedly, depending on how similar a competitor is, it may simply be confused with the target during retrieval (Diana et al., 2004; Schurgin et al., 2020).

Other interference perspectives focus on changes to the memory of the item itself. This could involve weakening of the memory representation of the item overall, or specific changes to features (see Interference resolution, below). Because the means of measuring memories at the feature level have been limited until recently, mechanistic accounts of changes to the memory themselves remain somewhat speculative in nature. Ultimately many of these potential interference mechanisms may occur depending on the circumstances, but determining when and to what degree they are occurring will require more research that fully maps the circumstances of interference and the effects thereof.

The role of item similarity in interference

Many factors go into the degree of memory interference experienced. One key factor is the degree of similarity between competing items. Since interference is triggered by similarity, it is intuitive to suspect that greater similarity between competing items leads to greater memory interference. In fact, there is quite a lot of research that supports this view from a broad array of paradigms (Anderson et al., 1994; Anderson & Spellman, 1995; Baddely, 1964; Baddeley & Dale, 1966; Chanales et al., 2017; Smith & Hunt, 2000; Watson & Lee, 2013; Yeung et al., 2013). While similarity does tend to cause interference, there are contexts where high similarity can be beneficial to memory in comparison to moderate similarity (Anderson, 2003; Bauml & Hartinger, 2002; Kahana et al., 2007; Lin & Luck, 2009; Mate & Baques, 2009; Sanocki & Sulman, 2011). This suggests that depending on the context, there may be a specific level of similarity where interference peaks and both lower and higher levels of similarity would tend to cause less interference.

This question has primarily been studied with dichotomous measures of memory, where forgetting an item is used as an indicator of greater interference. However, recently working memory studies have used continuous measures of memory features to address this question. This research suggests that highly similar items may damage the precision of remembered memory features, whereas less similar items may disrupt the ability to retrieve the memory at all (Li et al., 2020; Sun et al., 2017). These studies provide clues as to what the role of similarity in interference is, however it remains unknown to what extend these findings apply to long-term memory and what the neural mechanisms underlying this relationship is.

Neural origins of interference

One of the most influential frameworks for understanding the human memory system is the complimentary learning systems (CLS) framework (Battaglia et al., 2011; McClelland et al., 1995; Norman, 2010; Norman & O'Reilly, 2003; O'Reilly & McClelland, 1994; O'Reilly & Norman, 2002; O'Reilly & Rudy, 2001; Schapiro et al., 2017; Schlichting et al., 2015). Under this perspective, there are two basic types of learning: (1) a 'fast' system designed to remember specific events and (2) a 'slow' system for extracting generalities over time. Although all memories can still be impacted by the neural architecture of the slow system, this dissertation focuses on the fast system—which supports the formation of distinct memories with a large amount of detail and where interference is an obstacle to be overcome.

The formation of distinct memories is largely driven by the architecture of the hippocampus. When a novel stimulus or event is experienced, the hippocampus forms distinct representations in CA3. This automatic formation of a distinct representation is known as pattern separation (Bakker et al., 2008; Yassa & Stark, 2011). Computational models suggest that the sparse coding in the dentate gyrus drives the formation of these distinct representations independently of the content of the memory (Norman & O'Reilly, 2003; O'Reilly & Norman, 2002; Schapiro et al., 2017). These unique, orthogonalized neural representations decrease the likelihood of interference.

When attempting to retrieve a memory, a partial retrieval cue can lead to the reactivation of the full memory representation in CA3 (O'Reilly & Norman, 2002). Computational models suggest that the recurrent connections in CA3 lead to a process known as pattern completion where the activation pattern in this region will converge towards the full memory representation. CA3 first outputs to CA1, where memory features are likely represented. The reinstatement of the memory then continues to cascade out to the entorhinal cortex and to connected and distributed cortical regions (Xue, 2018).

Memory interference likely arises from two competing needs in the memory system: (1) to represent the features of memories that may have little distinction, while simultaneously (2) forming distinct pathways for correct retrieval (Colgin et al., 2008). This tradeoff appears to be addressed by the outlined hippocampal architecture where pattern separation automatically creates *distinct representations* in dentate gyrus and CA3, then during retrieval, memory features are represented in broad patterns across cortical regions. Yet, despite a neural architecture that seems, in part, designed to diminish memory interference, it still quite often occurs. Further, since pattern separation occurs *automatically* when a new item is experience, it can't explain interference resolution. This means that instances of overcoming interference must be explained by another, *experience-dependent* neural process.

Adaptive memory changes as a route to interference resolution

Adaptive memory errors and distortions

Our memory system is not akin to taking a photograph where a perfect representation is stored, instead there are systematic errors and distortions (Schacter et al., 2011). Memory is better understood as a reconstructive process. Evidence for this comes from systematic patterns in memory errors, and from the discovery of neural overlap between areas involved in memory retrieval and with imagining the future (Benoit & Schacter, 2015; Schacter & Madore, 2016). Memories are shaped by our already formed preconceptions about the world (Tompary & Thompson-Schill, 2021), changes in the current environment (Brunec et al., 2018; Zheng et al., 2022), attentional focus (Hutchinson et al., 2016; Swan et al., 2016), behavioral goals during both encoding (Long & Kuhl, 2018) and retrieval (Favilla et al., 2018; Mack et al., 2016), and *other memories* (Scotti et al., 2021).

There are a great number of well-established memory errors and distortions, many of which reflect adaptive attributes of our memory system. For example, when participants are presented with a list of semantically related words (e.g. “shell”, “omelet”, “yolk”, “frittata”, “scramble”), they will tend to develop a false memory for a related word not on the list (e.g. “egg”; Deese, 1959; Roediger & McDermott, 1995). Although “egg” was never seen it may be adaptive to falsely remember that word because it reflects the gist of what was experienced, thereby better allowing you to apply that experience in the future (Schacter et al., 2011). Another example is when information later learned about an already experienced event is incorporated into the memory for that event. This reflects an adaptive memory change where memories are flexible enough to incorporate new information, however it can show up as an error when false information is incorporated into the memory (Schacter et al., 2011). By identifying these patterns, we can begin to see how certain memory “errors” are only errors in respect to veridical memory, not flaws in how the memory system is working. What is made clear from this perspective is that both the form and precision of memory measurement, and an analysis method that minimizes assumptions about what a memory error is, are critical.

The view that memory errors can be adaptive is key to understanding how memories may change in response to interference. For example, the act of retrieval can lead to forgetting in the context of interference. In one of the main paradigms used to study this effect, participants study the association between an item and a semantic category (Anderson et al., 1994; Hulbert & Norman, 2015). Then in a retrieval practice phase, half of the items from half of the semantic categories are cued for retrieval. After several rounds of practice, all items are tested on retrieval based on the cue. Unsurprisingly, participants have a better memory for the items that are practiced (RP+) compared to control items where that category was not practiced (Nrp). The key, consistent finding from this paradigm is that unpracticed items from the practiced categories (RP-) are recalled *less* well compared to the control items (Nrp). This effect is known as retrieval induced forgetting (RIF; Anderson et al., 2000; Bauml, 2002).

This memory error is quite often adaptive (Bjork, 1989). For example, consider learning two different techniques for cooking scrambled eggs. There are several differences, but one key difference is that in J. Kenji Lopez-Alt's version you add salt at the beginning, but in Gordon Ramsay's version you add salt at the end. Later when you practice Lopez-Alt's version, it might be bad to accidentally recall Ramsay's admonishment to not add salt at the beginning. That is an example of a non-on/off memory error that could be caused by memory interference. Critically, in this case, if you are practicing one egg technique, forgetting the other technique is adaptive and increasingly likely with more experience practicing the recipe.

Interference resolution

Forgetting is not always adaptive though; sometimes you want to remember both of the competitive items. Fortunately, this sort of memory interference can be overcome. In fact, overcoming this sort of interference can actually strengthen the once forgotten items (Storm et al., 2008). This was demonstrated by interleaving a relearning phase with the retrieval practice phase in a RIF paradigm. They found the typical RIF effect for RP- items that were not relearned. However, this forgetting actually enhanced learning for those items. In particular, for RP- items that were relearned, memory was better than for control items which were studied the same number of times. This suggests that the memory change involved in RIF is more than simply strengthening or weakening memory signals or retrieval routes; there is likely something more complex going on that works adaptively to prioritize certain memories when circumstances suggest that is beneficial, but also allow for correct retrieval of two or more competing memories when that is beneficial.

A neural network model that would explain this effect comes from Hulbert and Norman (2015). They model a memory as a combination of its constituent features, with each feature as a node in the network (Fig. 1.1). This means that when items are competitive with one another due to similar features, many of the nodes may be overlapping. Under this model, it is the representational overlap which causes interference. In a RIF paradigm, when one item is retrieved, the features of that memory are activated and the connections between those nodes become strengthened. This enhanced memory strength increases the likelihood that the RP+ item will be correctly remembered latter. At the same time, because there are shared features with the completive RP- item, that item also becomes weakly reactivated. This weak activation works to weaken the connection between the unique features of the RP- item and the features shared with the RP+ item. This process decreases the likelihood that the RP- item will be correctly recalled later. This is especially true if the cue is related to the shared features—another retrieval route may be needed. This explains the typical RIF effect. If the RP- item is then relearned, the unique features are strongly activated and other features that were not a strong part of the original representation may become emphasized more. This activation strengthens the connection between this new constellation of nodes. Over time, through interleaved practice, both memories can become strong, but with more distinct neural representations. This process of separating the memory representations over the course of learning is known as pattern differentiation or repulsion.

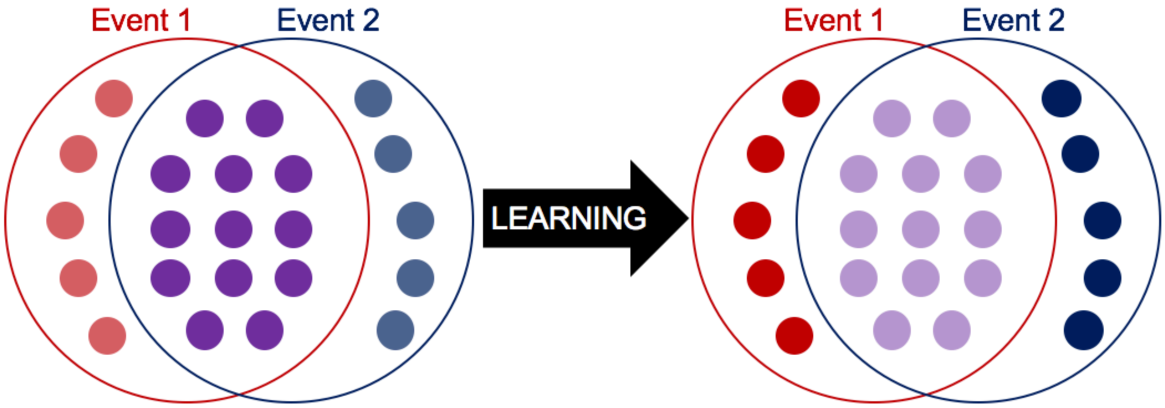

Figure 1.1. Simplified summary of the Hulbert and Norman (2015) model. Left. The memory of an event can be thought of as constellation of individual features (circles). When two events are similar, they likely have many shared features (purple) and may interfere with one another. Right. Over the course of interleaved learning, the model predicts that the unique features of each event (red: event 1, blue: event 2) become strengthened and the shared features become weakened.

Although the Hulbert and Norman (2015) model makes sense from a theoretical or computational perspective, there is limited experimental evidence for all aspects of their account. They predicted that repulsion would be detected in the form of lower similarity in hippocampal representational patterns. In their study, they found that greater repulsion in left hippocampus was associated with greater learning on the RP- items relative to control items. Thus they established initial evidence for a relationship between repulsion and learning. Importantly, they had no measurement of memory features, thus no way of evaluating that aspect of their model.

Others have expanded on this initial evidence. One study that used an associative learning task found that the hippocampal representations of similar scene images were driven apart (repulsion) through learning (Favila et al., 2016). Critically, these neural changes were associated with interference reduction. Another study tracked hippocampal representations as participants learned routes (Chanales et al., 2017). The key manipulation was that each route had an overlapping portion of the route with one other route, making those two memories competitive with one another. They found that prior to learning, the neural representation in the hippocampus of overlapping routes were as similar to one another as non-overlapping routes. This makes sense according to a pattern separation account (see Neural origins of interference, above). Critically, the hippocampal representation of overlapping routes became more dissimilar over the course of learning, consistent with a repulsion account. This effect was specific to more difficult trials, i.e. where interference needed to be overcome. Both of these studies could be explained by the Hulbert and Norman (2015) model, but without reference to specific features, other explanations cannot be ruled out.

More recently, one study looked at pattern similarity changes between competitive scene images over the course of learning (Wanjia et al., 2021). They found that the repulsion effect was specific to CA3/dentate gyrus subfield (as opposed to CA1), consistent with theories of hippocampal function (see Neural origins of interference, above). As evidence of the adaptive impact of these changes, the effect occurred specifically when interference was overcome and was strongest for items with the greatest initial interference. As an initial way to link repulsion to memory features, the repulsed activity patterns carried relatively more information about the correct compared to the incorrect learned association. Another recent study has also shown evidence for a shift in information after interference resolution (Zhao et al., 2021). Here they focused on changes in cortical regions and identified regions where greater information on the unique feature was represented for competitive items. They did not find evidence for overall repulsion, but these results are consistent with the idea of interference resolution involving shifts in the representation of specific memory features.

Overall there is strong and growing evidence that interference resolution, at least in part, involves a shift in neural representations such that competitive items are further apart in representational space. The extent to which this shift is consistent with ideas of Hulbert and Norman (2015) remain unresolved. One alternative that could potentially explain these results is the idea of hippocampal remapping. It has long been documented in animal studies that there are place cells in the hippocampus that preferentially fire in certain locations. When these animals are put in a new context, or believe they are in a new context, there is a rapid “remapping” where cells fire for new preferred locations (Bostock et al., 1991; Colgin, et al., 2008; Muller & Kubie, 1987; Wills et al., 2005). Although the evidence for this phenomenon is strongest in navigation, there is growing evidence that the same logic could apply to anything that the hippocampus is interested in (Colgin, et al., 2008; Wanjia et al., 2021). Thus, another way interference could be resolved is through the association of distinct internal contexts with competitive stimuli. That is, these neural changes in hippocampus could be unrelated to feature changes.

Thus, there are multiple potential models that are consistent with the idea of neural repulsion—and that is focusing specifically on the hippocampus, when other regions are likely playing a role as well (Zhao et al., 2021). In order to provide evidence that these neural changes are related to memory features, we must establish whether these neural changes are having an impact on *how* the item is remembered (not just if). Further, we need a way to decode these pattern similarity shifts into a meaningful feature space that corresponds to how the memory themselves may change.

Measuring the feature space of memory

Behavioral measures of memory content

Most long term memory studies have focused on whether an entire event is remembered or forgotten (Cooper & Ritchey, 2019). However, memory is not an all-or-none event—memories can be measured in much more informative ways. For example, a recalled item can be measured along a self-report rating scale in terms of how confident they are in their retrieval or how vividly it was retrieved (Kuhl & Chun, 2014; St-Laurent et al, 2015; Bonnici et al., 2016; Ford & Kensinger, 2016). Measuring perceived vividness captures the idea that memories can be recalled at a more gist-level, or with much greater detail (Brady et al., 2008; Schacter et al., 2011). What perceived vividness fails to capture is that the individual features of a memory can vary in vividness and may be distorted. Memories are best understood as a multi-dimensional constellation of features (Cooper & Ritchey, 2019; Horner & Burgess, 2013; Horner & Burgess, 2014; Xue, 2018). Therefore, researchers have increasingly begun to measure long-term memories along continuous feature dimensions, probing features such as, location on the screen (Berens et al., 2020; Harlow & Yonelinas, 2016; Nilakantan et al., 2018), orientation (Richter et al., 2016), and color (Brady et al., 2013; Chanales et al., 2020).

Continuous measures of feature memory are tremendously useful because they can be utilized to estimate both the precision and accuracy of individual memory features. I define precision as a measure of how detailed the memory for a particular feature is. For example, you could remember that a person's eyes are blue, or you could have a more precise memory for a specific shade of blue. Importantly, although precision is often conflated with accuracy, I view the two as independent. That is, you could have a very detailed memory for the eye color and be wrong. I define accuracy as how close to the true value a feature memory is.

There are multiple ways to measure precision and bias, one of the most straightforward ways is to bin the data. You might create some range around the true value that allows some small degree of error to be considered accurate, then you can create bins further away to indicate some degree of inaccuracy (e.g. Nilakantan et al., 2017). Another similar approach would be to take the absolute value, as a measure of the average distance from the true value. An alternative approach, mixture modeling, views responses as mixture of multiple underlying distributions (Zhang & Luck, 2008). For example, the overall distribution can be driven by some responses that reflect a successfully retrieved memory with some amount of precision and some responses that reflect random guessing. This analysis is helpful for determining not only the precision and accuracy of memory features, but also how often they are retrieved at all.

Regardless of which analysis approach is taken, in the context of interference, we might expect for there to be distortions in feature memory. Therefore, I view not only a measure of accuracy as important, but a measure of whether there is any directional bias in errors made. Again bias, can be independent of precision. Traditionally in studies of memory interference, we might expect the confusion caused by two competing memories to lead to integration, where there are no longer two distinct memories and only a more gist-level recollection of both. This would cause memories to be recalled with a bias *towards* each other. The Hulbert and Norman (2015) model, however, suggests that over the course of overcoming interference, the differences between competing items becomes highlighted. Under these circumstances we might see a bias *away* from a competing item.

Measures of precision, accuracy, and bias, have been widely used to study the impact of interference in working memory (e.g. Sun et al., 2017). However, they have seldom been used to study *long-term* memory interference. There are important properties the stimuli need in order to apply this approach. (1) In order to test the predictions of the Hulbert and Norman (2015) model, the stimuli need at least two independently measurable dimensions that can each be manipulated to act as a shared or unique feature. (2) These dimensions need to be perceptually-important. For example, two tree images can cause interference with one another and an experimenter could create an underlying dimension that defines that perceptual difference. However, that dimension would not be meaningful to participants based on viewing those two images alone. Therefore, you could not expect to detect changes to memory that align with this latent dimension. In contrast, for perceptually-important dimensions (e.g. color or location) we would expect to be able to detect changes in memory. (3) Interference needs to be restricted to where the experimenter intends it to occur. In working memory studies of interference, trials can be treated independently because information does not need to be retained beyond that trial. In contrast, in long-term memory studies the information needs to be retained throughout the experiment. Thus, for example, utilizing the angle of a gradient as a continuous memory measure works when each trial can be treated independently. However, in a long-term memory study, the gradient stimuli would all interfere with one another.

In Chapter 2 I develop a method that allows for the creation of synthetic face stimuli that meet these criteria. In Chapter 3 I demonstrate an approach to analyzing memory feature data in a way that decouples accuracy and precision. With the view that memory distortions often serve an adaptive function, I link both the accuracy-independent measure of precision and bias for a specific feature with improved performance on a separate measure of memory.

Neural measures of memory content

Episodic memories are supported by a broad pattern across many cortical regions. Successful retrieval of those memories is associated with reactivating similar neural patterns as were originally elicited by the event (Kuhl et al., 2011; Xue et al., 2010; Zeithamova et al., 2012). Recent advances in fMRI data analysis have improved the ability to characterize neural representations in terms of the specific information they are representing (Cohen et al., 2017; Davis & Poldrack, 2013; Norman, Polyn, et al., 2006; Rugg et al., 2002). This can be helpful to determine differences between brain regions, and of particular interest here, how information is transformed over the course of processing and perhaps over the course of time through learning.

Multiple fMRI analysis approaches can to some degree measure the content of memories, including univariate activation and adaptation (Davis & Poldrack, 2013). Of most interest here are multi-voxel pattern analysis (MVPA) techniques that take into account not only activation levels of individual voxels or regions, but are instead driven by distributed representational patterns (Kriegeskorte et al., 2008). These approaches tend to have the greatest sensitivity, particularly when attempting to delineate representations in multi-dimensional feature spaces (Davis & Poldrack, 2013).

One approach that has proved particularly effective in this domain is representational similarity analysis (RSA). RSA involves creating a dissimilarity matrix between pairs of stimuli or conditions based on neural activity within a brain region (Kriegeskorte, et al., 2008). This approach utilizes all informational content available and maps it to a common representational space that can be compared across region or time, or compared to stimulus feature spaces. In condition-rich designs where there are many unique stimuli, this approach can be very effective at mapping a complex representational space (Drucker & Aguirre, 2009; Kriegeskorte et al., 2008; Nestor et al., 2016). In experimental contexts where a small number of competitive stimuli need to be learned, however, it may be more difficult to make that type of mapping.

A particular form of this approach has recently been used to focus on the similarity of competitive pairs (rather than a full feature space), tracked over the course of learning (Chanales et al., 2017; Wanjia et al., 2021). RSA or other approaches (e.g. Chadwick et al., 2011) that distinguish items without reference to specific features are adept at distinguishing similar items in the hippocampus. Techniques such as multi-dimensional scaling (MDS) can further help to visualize and interpret representational changes in the form of a feature space. However, these MDS features do not necessarily correspond to interpretable feature memory changes. Further, the ability to distinguish competitive items in other cortical regions—where these representations are shifted to in the long-term and where the feature information is reactivated—may be more limited with this approach.

The MVPA approach that this dissertation focuses on is decoding. This approach puts the output into a meaningful dimension that can correspond to hypotheses about how memory features change in response to interference. Decoding approaches have the potential to be more powerful because they do not weigh all voxels equally (as RSA does). Although decoding may have first appeared to have limited power (Carlson et al., 2003; Cox & Savoy, 2003; Haxby et al., 2001), the upper limits in power and specificity have continuously been pushed (Huth et al., 2016; Mozafari et al., 2020; VanRullen & Reddy, 2019). This has included increasingly complex output from complex stimulus classes (Dado et al., 2022; Lin & Hsieh, 2022). Thus, this approach has the tantalizing potential to bring meaning to shifts in overall neural pattern similarity.

Decoding could be a powerful approach to bring meaning to the shifts detected through similarity based approaches that have been used to find repulsion in the context of memory interference (see Interference resolution, above). If these similarity shifts are driven by changes in feature information, then these shifts could be decodable. Similarity based approaches have been demonstrated to detect small but meaningful shifts in neural representations over time; it is unclear whether decoding will be able to reliably measure small shifts, but if it could, the implications would be quite powerful. In Chapter 2 I develop a set of dimensions that describe a large set of stimuli that are good candidates to be used in the study of memory interference (faces). In Chapter 4 I describe an approach to decode those dimensions from fMRI activity.

Goal and structure of the dissertation

The primary goal of the dissertation is to develop and validate an approach to studying how the *features* of memories change in response to interference. In Chapter 2 I will focus on the development and validation of the methodology. In Chapter 3 and 4 I will focus on specific applications of this methodology. I will conclude in Chapter 5 with a discussion of how this approach can be leveraged going forward in the context of long-term memory interference and in cognitive neuroscience more generally.

In Chapter 2, I document the development and initial validation of a set of face stimuli standardized for the use in psychology experiments, with a number of useful metrics and the ability to control and manipulate. Face stimuli are a perfect stimulus class for use in long-term memory studies with high interference because humans are experts at processing faces and can later remember fine-grained differences as having distinct identities. Further, as established in this chapter, multiple dimensions at multiple levels of neural processing can be experimentally controlled and probed from memory.

I follow that in Chapter 3 with a demonstration of this methodology in a high interference behavioral setup. The face stimuli are used to create competitive pairs of stimuli that are matched in all features except for one, where they are only slightly different. The experiment tracks associative memory over the course of experiencing and then overcoming memory interference. The face methodology is then utilized again to probe feature memory along the same dimensions the stimuli were manipulated along. I found that feature memories that are diagnostic of the difference between competitive items are both biased and recalled with greater precision in response to interference. Further, I found these memory changes to be adaptive for learning.

In order to eventually apply this methodology to decoding neural representations in the context of interference, we first need to establish the ability to decode and to identify the most decodable features. Thus, in Chapter 4, I investigate the ability to decode face features from perceptual data. I discuss differences in the ability to decode different data-driven face components and subjective ratings, and how those differ between brain regions.

I conclude in Chapter 5 with a summary of our results and the broader implications of those findings. I also discuss how the approaches applied in Chapters 3 and 4 can be utilized in concert moving forward. The goal is to help bridge the gap between neural and behavioral perspectives on resolving memory interference. A convergent approach is key to understanding how memory change in response to interference supports the vast human memory capacity.

Chapter II

STANDARDIZED SET OF 1,148 FACE STIMULI WITH LANDMARKS, SORTING, AND RATING DATA

This chapter contains unpublished co-authored material. Maxwell L. Drascher is the primary author of this chapter with input from his advisor Brice A. Kuhl. Drascher and Kuhl designed the study together. Drascher conducted all data collection, and wrote the scripts for experiment presentation, data analysis, and figure creation. Drascher wrote the manuscript with editorial assistance from Kuhl.

Introduction

Face images represent a unique category of visual stimuli given the fact that they are relatively uniform and contain common features, but humans can still perceive and later remember subtle differences. Human expertise in faces also makes those subtle differences measurable and amenable to parameterization. These properties make faces an appealing class of stimuli that can be leveraged to study a broad range of cognitive domains. Developing the ability to measure and manipulate faces along distinct dimensions is key to unlocking the full potential of face stimuli in experimental settings.

Faces are comprised of many features, including measurable physical dimensions such as eye color and skin tone. Faces are also comprised of a variety of high-level, socially-relevant dimensions such as gender, trustworthiness, and dominance. Even these seemingly more abstract dimensions are perceived similarly across different participants, which makes them measurable (Oosterhof & Todorov, 2008). However, the ability to tightly control and reliably measure higher-level face dimensions is not as straightforward as many of the most commonly used feature spaces in cognitive research, such as color (e.g. Bays et al., 2009; Chanales et al., 2021; Zhao et al., 2021; Zhang & Luck, 2008), orientation (e.g. Haynes & Rees, 2005; Kamitani

& Tong, 2005; Korkki et al., 2020; Pertzov et al., 2017), or location on the screen (e.g. Berens et al., 2020; Harlow & Yonelinas, 2016; Nilakantan et al., 2017).

One approach to manipulating faces is to use actors showing different facial expressions (e.g. Benda & Scherf, 2020; Chung et al., 2019; Conley et al., 2018; Ebner et al., 2010; Engell & Haxby, 2007; Furl et al., 2013; Said et al., 2010; Thomas et al., 2001; Tottenham et al., 2009). This approach does not directly create a continuous feature space though. One method of generating a continuous face space is morphing, which utilizes two or more specific face images to generate a continuous space between them (Steyvers, 1999; Leopold et al., 2001). This technique can generate a continuous dimension between two different emotional expressions of the same face (e.g. Arsalidou et al., 2011; LaBar et al., 2003; Sato et al., 2004; Won et al., 2020), but this space may not align well with the true space between those emotions (Hays et al., 2020). Morphing can also generate a dimension between any specific face images. This type of dimension is a great tool in experimental designs where it can be implicitly learned during an experiment, for example in a category learning paradigm (e.g. Ashby et al., 2020; Goldstone et al., 2001). However, it is less useful at generating perceptually-important dimensions (e.g. affect).

Alternative approaches to parameterizing faces include approaches that generate dimensions based on the variance in base image properties across a large pool of faces. One early approach to this was eigenfaces, which are generated from a principal component analysis (PCA) on the pixel intensity values across three color channels (Turk & Pentland, 1991). This approach is powerful given the high degree of information it can capture with a limited number of components. However, because the components are completely data-driven, the individual components are highly influenced by the stimuli used, are difficult to interpret, and are often poorly aligned with features important for face perception. This approach was subsequently improved with the active appearance model (AAM; Chang & Tsao, 2017; Cootes et al., 2001; Edwards et al., 1998). The AAM is more labor-intensive, with the requirement of landmarking the face stimuli, but yields greater reproduction of the face images with fewer components. This approach also yields components that do not necessarily align with human perception of faces, however they tend to be more interpretable both individually and with the inclusion of two broad component groupings (shape and appearance). Both eigenfaces and AAM create measurable and manipulable face dimensions, however they lack a clear, innate connection to features important to face perception.

An alternative approach is to use artificially generated and manipulable face images (e.g. Roesch et al., 2011). For a long time, the capabilities of these types of stimuli were limited and were often unavailable to researchers, however the technology available is becoming increasingly realistic (e.g. Hays et al., 2020; Peterson et al., 2022). For the tightest experimental control, one of these options may be optimal. However, there is evidence that synthetic faces may be processed differently than real faces (Balas & Pacella, 2015; Schindler et al., 207; Wheatley et al., 2011). Thus, although synthetic images are extremely valuable, they should be used with caution, especially if you want to study face processing specifically. Until it is demonstrated that there are no differences in behavioral and neural responses, real face stimuli will remain a valuable resource for cognitive scientists.

Researchers looking to use real face stimuli in their experiments have many options to choose from (e.g. Bainbridge et al., 2013; Benda & Scherf, 2020; DeBruine et al., 2017; Ebner et al., 2010; Ma et al., 2015; Minear & Park, 2004; Walker et al., 2018). Although many of these databases may have lacked diversity before, they are becoming increasingly diverse (e.g. Chen et al., 2021; Chung et al., 2019; Conley et al., 2018; Lakshmi et al., 2020; Ma & Wittenbrink, 2020). All of these databases contain information on certain stimulus properties, however, depending on the needs of a particular experiment, the list of compatible databases may be narrowed significantly or not exist at all. This is particularly true when there are multiple, complementary purposes for the stimuli.

One important metric that is not often available for face stimuli is an overall measure of similarity. Similarity is a key measure in many experimental designs, however it is difficult to apply to more complex stimulus classes because similarity on any one dimension or combination of dimensions does not necessarily correspond to overall perceptions of similarity (Jiang et al., 2021). Therefore, data specifically meant to capture overall similarity, such as sorting faces into groupings, is required. Another important metric that is often not available is facial landmarking. When, looking to apply the AAM, the list of face database options is either limited substantially (e.g. Koestinger et al, 2011; Milborrow et al., 2010) or requires the labor-intensive process of manually landmarking a new stimulus set. Although properties like these are available in some circumstances, when selecting face stimuli there are cases where it would be beneficial to use the same stimuli for many purposes. Thus, a large database that contains information on not only subjective ratings, but also less commonly available information such as sorting data and landmark positions, may be the optimal option.

The current manuscript describes the data collection and validation process for a broad set of data that describes face stimulus properties and offers advice for future applications based both on previously published uses (Drascher & Kuhl, 2022) as well as potential future uses. The ultimate goal is to create a freely accessible resource of face stimuli with data available that facilitates their use across a broad array of purposes. The database contains a total of 1,148 faces, all forward facing, and cropped to a uniform size and position in the frame. The faces were selected to be diverse in terms of gender, age, ethnicity, and facial expression. The key distinguishing features of this corpus are the breadth and uniqueness of data available on this size of a face corpus. All images have been independently rated on several important social dimensions, have been sorted based on appearance, and have been hand landmarked.

This diversity of information allows for the use of face stimuli with a high degree of experimental control on reliable, perceptually-important dimensions (Drascher & Kuhl, 2022), while being large enough to be used as a training set in neuroimaging-based, image reconstruction designs (Lee & Kuhl, 2016). This dramatically increases the utility of the face stimuli by allowing the same stimuli to be used with multiple potential applications and facilitating the bridging of behavioral and neuroimaging findings.

Methods

Face image corpus

A total of 1,148 faces were selected from a variety of online sources (see Lee & Kuhl, 2016; a small number of faces [8] were removed from the 1,156 in this set for having attributes highly distinct from the rest of the set and were thus unlikely to be successfully reconstructed). All faces were forward-facing and cropped and resized to 179 x 251 pixels. The faces were selected to be diverse in terms of gender, age, ethnicity, and facial expression. The full corpus is available at: https://osf.io/4uydh.

Face image landmarking

All of the face stimuli were hand landmarked in each of 62 locations (see Fig. 2.1 for an example; see Chang and Tsao (2017) for a similar landmarking scheme). Landmark locations were chosen to represent the overall shape of the face as well as the shape and relative position of internal features. The landmarks locations share a lot of overlap between what was used previously (Chang & Tsao, 2017), however there are many differences that account for differences in the stimulus set. For example, in our scheme, no landmarks were included to track the top of the head because the face images were cropped there, however landmarks were included to mark the hair line. In our scheme, we also included a high number of landmarks to track eyebrow and mouth shape, in order to capture the variance in expression in this set. In total, 9 landmarks tracked the cheek and jaw line, 10 tracked eye shape, 12 tracked eyebrow shape and position, 16 tracked mouth/lip shape, 11 tracked the nose shape, and 3 tracked the hairline. Most critically, after initially creating the landmarks, the positions were adjusted through piloting in order to capture the variance within the stimulus set, while also being able to be consistently applied across stimuli and different raters.

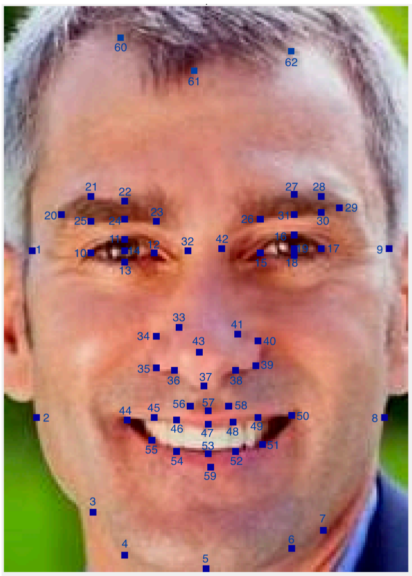

Figure 2.1. Example of the 62 landmark positions on one of the face images. The positions were designed to capture the variance in the shape of faces in the corpus (see https://osf.io/4uydh/). This landmarking was completed on 1,148 unique images.

The consistency of landmark positions was measured with a series of two-way mixed-effects, agreement intraclass correlation coefficients (ICCs) on the vertical and horizontal position of each landmark (124 total). The two-way mixed-effects model treats the items as random, but the effect of rater as fixed. We opted for that because we were not interested in generalizing our findings to other potential raters. Agreement ICCs were used because the consistency of the absolute position of the landmarks is critical. After piloting, the majority of landmarks had high reliability within and between raters (majority of landmarks above 0.75). A small number of landmarks maintained low ICC (below 0.4), however we attribute this to the landmark locations having low variability between images (due to properties that were standardized in the set such as face position), which led to poor reliability as assessed by ICC, but small absolute differences between raters.

Eight research assistants were then trained on the landmarks of one of the original two raters. During training, performance was assessed with a series of ICCs compared to the original rater and/or other raters that had already completed training. This helped pinpoint landmark locations that needed to be fine-tuned for each rater. Training lasted until the ICC was high (consistently above .75 for most landmarks, excluding low-reliability landmarks explained above).

Six research assistants (out of the eight) completed training with high reliability and collectively landmarked the remaining images. In order to continually evaluate performance and consistency between raters, images were periodically repeated across different raters. In total, 510 images were landmarked by two or more raters and 36 were landmarked by 3 or more. This allowed for the continued evaluation of reliability. In cases where large differences were identified between raters (greater than 5 SDs on one landmark), visual inspection revealed that in the vast majority of cases the large differences reflected an ambiguous property of the stimulus (e.g. whether dark pixels represent a shadow or the continuation of an eyebrow), rather than an error.

After landmarking was completed, we made small automatic adjustments for certain landmark positions. Specifically, landmarks that fell near the edge (e.g. along the jawline) were sometimes placed just outside the image range, those landmarks were automatically shifted back within the image. Additionally, in order to handle landmarks for the bottom lip accidentally being place slightly above the top lip (in cases where the image had a closed lip), those landmarks were shifted 2 pixels apart vertically.

In instances where stimuli were landmarked by multiple raters, we used the average landmark position after this initial preprocessing. A small number of stimuli (54) were landmarked multiple times by one rater. These repeated landmarks were averaged together prior to averaging with landmark positions from other raters.

Active appearance model application

Application of the active appearance model (AAM) was similar to prior approaches (Chang & Tsao, 2017; Cootes et al., 2001; Edwards et al.,1998; Van Ginneken et al., 2002), however the approach needed to be modified for application to color images. First a principal component analysis (PCA) was run on the vertical and horizontal positions of the 62 landmarks across all 1,148 images. This generated shape components and the mean face shape. Many of these components likely reflect noise, thus we filtered out the bottom 1% of components in terms of variance explained, leaving 61 shape components. In order to generate appearance components, each stimulus was smoothly warped using inverse weighted interpolation (with a radius of 10 and power of 5) to match the mean face shape (Bookstein, 1989). Warping of the images was applied in MATLAB (adapted from: Archibald, 2009). The process of warping the images to the mean shape created a set of shape-free face stimuli. A PCA was then performed on red/green/blue intensities across all pixels (179 x 251) of the shape-free faces. This generated a set 1,148 appearance components, however we retained the components that explained 99% of the variance, leaving 753 appearance components. The AAM was applied in MATLAB using a modified version of the Active Shape Model (ASM) and Active Appearance Model (AAM) package (Kroon, 2012).

Subjective similarity sorting

Procedure. This data was collected on a slightly larger sample of face stimuli, prior to the removal of 6 images (see Face image corpus). A group of 6 research assistants completed this task. On each trial, a random sample of 65 face stimuli were presented on the screen simultaneously (Fig. 2.2). Sorters were instructed to group the stimuli based on “which faces looked more closely genetically related to one another (i.e. the faces grouped together are more likely to have common ancestors).” We used this language so that similarities or differences due to gender, age, and hairstyle would play a minimal role in the sorting. Faces were place into groups by clicking the mouse to put a square of a certain color around that face. Each color was associated with a number on the keyboard (0-9), which allowed the sorters to change the color/group. There were no restrictions on how many stimuli needed to be in a group, however every face needed to be put into a group, even if it was by itself. With each new set of stimuli, sorters could group the faces into up to ten groups, but were not required to use all ten. The group numbers were treated as independent on each trial, so if a stimulus was in group 1 on one trial, that could be ignored on subsequent trials. The only relevant information was which stimuli were grouped together. There was no time limit or restriction on the ability to switch groupings. Pressing the return key sorted the faces into the groups visually, providing the sorters the chance to make any changes. When they were finished, they pressed space to proceed to the next trial.

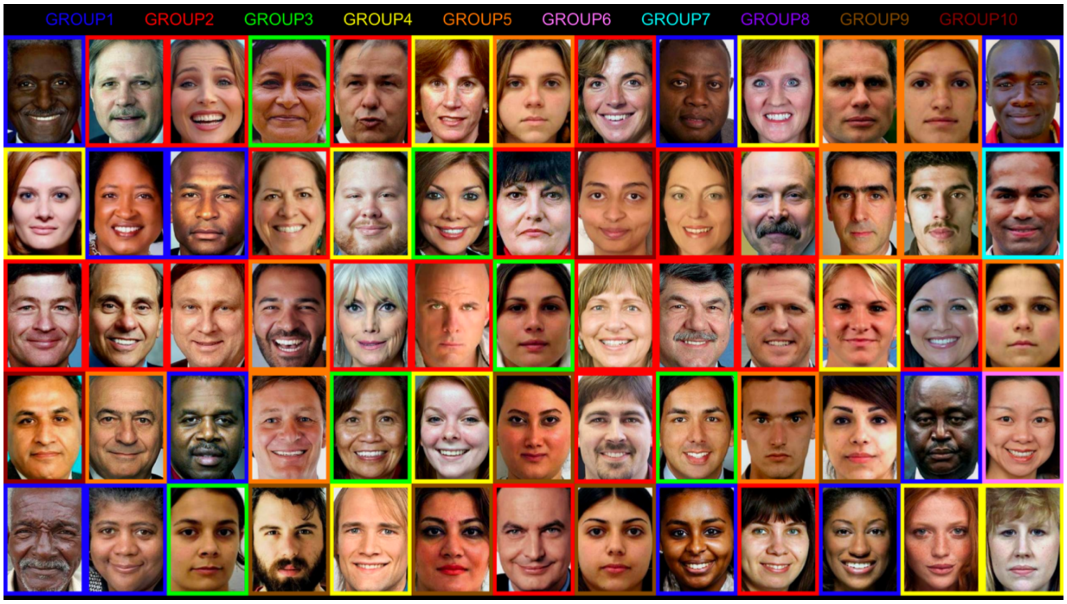

Figure 2.2. Example trial from sorting task. A total of 65 images were presented on the screen at a time. Participants used their mouse to click on an image to put a colored box around an image, with the color indicating one of ten possible groups (top). The color was switched by pressing the corresponding number on the keyboard. Participants were required to sort every face into a group, with no restrictions on how many groups were used or the size of the group.

Faces were presented in pseudo-random sets, where across all sorters, every face was presented with every other face at least once. This design allowed for the creation of a dissimilarity matrix (collapsed across sorters) that had information from at least one trial for every pairwise combination in the set (1,335,180 combinations). The overall order of the trials was random, as well as the position of the images on the screen. The images were displayed at full size (251 x 179 pixels) in 13 columns and 5 rows on a 27-inch iMac screen.

Reliability analysis. Using the sorting data collapsed across all sorters, we generated a dissimilarity matrix across all stimuli with each calculated as 1 minus the percentage of times each pair of stimuli were grouped together when they appeared in the same trial. In order to measure the reliability across all raters, we generated dissimilarity matrices based on every combination of 3 sorters (a total of 20 dissimilarity matrices). For each dissimilarity matrix in this set, there was one corresponding matrix with a set of 3 different sorters. We then calculated the correlation between all 10 non-overlapping pairs of matrices.

Subjective ratings

Participants. Face ratings were collected online via Amazon Mechanical Turk (MTurk). A total of 111 MTurk participants completed the rating task. All participants were located in the United Sates, were at least 18 years old, and had a MTurk job approval rate of 0.9 or above. Informed consent was obtained in accordance with procedures approved by the University of Oregon Institutional Review Board. Participants were paid $2.50 for completion of their ratings. We set a goal of obtaining 5 unique raters for each stimulus and rating type, a total of 100 participants (each participant rated 25% of the stimuli). Participants were removed based on a set of exclusion criteria that ensured compliance and effort in the task (see below). Participants that were excluded from further analysis were replaced until we reached this recruitment goal.

A total of 11 participants were removed based on our exclusion criteria. We excluded participants who responded too quickly (less than 500 ms) on a high percentage (greater than

15%) of trials. This was intended to exclude participants who were not engaging with the task and were just clicking quickly to get to the next trial; two participants were removed based on that criteria. Additionally, we excluded participants who were inconsistent in their ratings of repeated stimuli ($r < 0.4$); nine participants were removed based on that standard. The variability in responses was individually examined for each participant, to ensure that there was variability in the responses made and that there were no systematic patterns in the responses (e.g. repeatedly clicking the same number in consecutive trials). No additional participants were removed based on systematic response patterns.

Procedure. On each trial, participants were presented with one face stimulus in the center of the screen. Below each stimulus were nine buttons representing the range of the rating scale. Participants were instructed to rate each face based on their personal opinion by using their mouse to click the corresponding number (1-9) on the screen. Each participant made ratings on one of five dimensions: dominance, trustworthiness, attractiveness, happiness, or masculinity/femininity. In the first four cases the prompt was, “how dominant/trustworthy/attractive/happy is this this person?”, with 1 labeled as “not at all” and 9 labeled as “extremely”. For masculinity/femininity, 1 was labeled as “extremely feminine” and 9 was labeled as “extremely masculine”.

This data was collected on the same slightly larger set of faces as the sorting task. Each participant was randomly assigned to one of four possible stimulus pools, containing 25% of the stimuli. Participants were presented with a total of 289 unique stimuli, with 10% (29) randomly repeated for a total of 318 trials.

Reliability analysis. Responses made quicker than 500 ms were presumed to be errant and were removed from all analyses, this occurred for 23/100 participants (excluding participants already removed), but very few times ($M = 5.30 \pm 10.86$). We assessed reliability both within and across participants. Reliability within participants was measured with a correlation between repeated images and with an average of the absolute differences. We then averaged these results for each rating type. In order to assess the inter-rater reliability of these ratings, we calculated a two-way random-effects ICC on the consistency amongst raters. Here, we did not use agreement ICCs as we did with the landmark positions, because we intended to standardize the ratings, so the absolute rating number was not important, only the relative positions of face stimuli. We ran this analysis separately for each rating type and stimulus pool, and then averaged the findings for each rating type. Missing data was omitted in a listwise way.

Results

Landmark validation

As an initial test of reliability, one rater landmarked 54 images twice. As a measure of test-retest reliability, we ran a series of two-way mixed-effect, agreement ICCs on the vertical and horizontal position of each landmark (124 total) for the repeated images. The ICC was high across all landmark positions (*M* = 0.93 ± 0.091, *Median* = 0.95, range: 0.18-0.99; 99.2% above .4), indicating high reliability for this rater.

As a measure of the reliability of the landmarks across raters, we used the same ICC analysis but with different raters rather than repeated images from the same rater. The stimuli which overlapped between raters differed, so we calculated the ICC separately for every pairwise combination of raters who had overlap in stimuli landmarked. Out of 21 possible pairwise combinations between raters, there were 15 combinations with overlapping images, with each rater included at least twice. Among the 15 combinations, there was a lot of variability in the number of overlapping stimuli (*M* = 102.33 ± 92.68, *Median* = 79, range: 3-305). For all pairwise set of raters we calculated the mean across all 124 ICCs. The ICC was consistently high (*M* = 0.78 ± 0.06, range: 0.68-0.87). We also calculated the percentage of ICCs within each pair that was above 0.4, with the percentage consistently high (*M* = 92.0% ± 6.2%, range: 80.7-

99.2%), suggesting that for the vast majority of landmarks, there was no concern about reliability. In fact, we calculated the same statistic for the percentage of ICCs above 0.75, and found that the majority of positions were reliably high (*M* = 71.1% ± 10.4%, range: 51.6-87.1%). Although there were a small number of landmark positions that had low ICCs, these were the landmarks identified during development that had low variance across the stimuli in the set (see Methods and Discussion).

Active appearance model application

Previous uses of the AAM have utilized the top 25 shape and top 25 appearance components, because those components alone capture the majority of the visual variance (Chang & Tsao, 2017). In this instance, the top 25 shape components capture 95% of the variance in landmark position, and the top 25 appearance components collectively capture 79% of the shape-free visual variance (see Fig. 2.3,4 for a visual representation of the top components). Thus, in this analysis, 50 components explain most of the variance and act as a potentially good cut-point for maximizing efficiency in representing face images. However, depending on the purposes of utilizing the components, there may be other cut-points that make sense. We found that visual reconstructions of face images were strong prior to 50 components, but also continued to improve up to and beyond that point (Fig. 2.5).

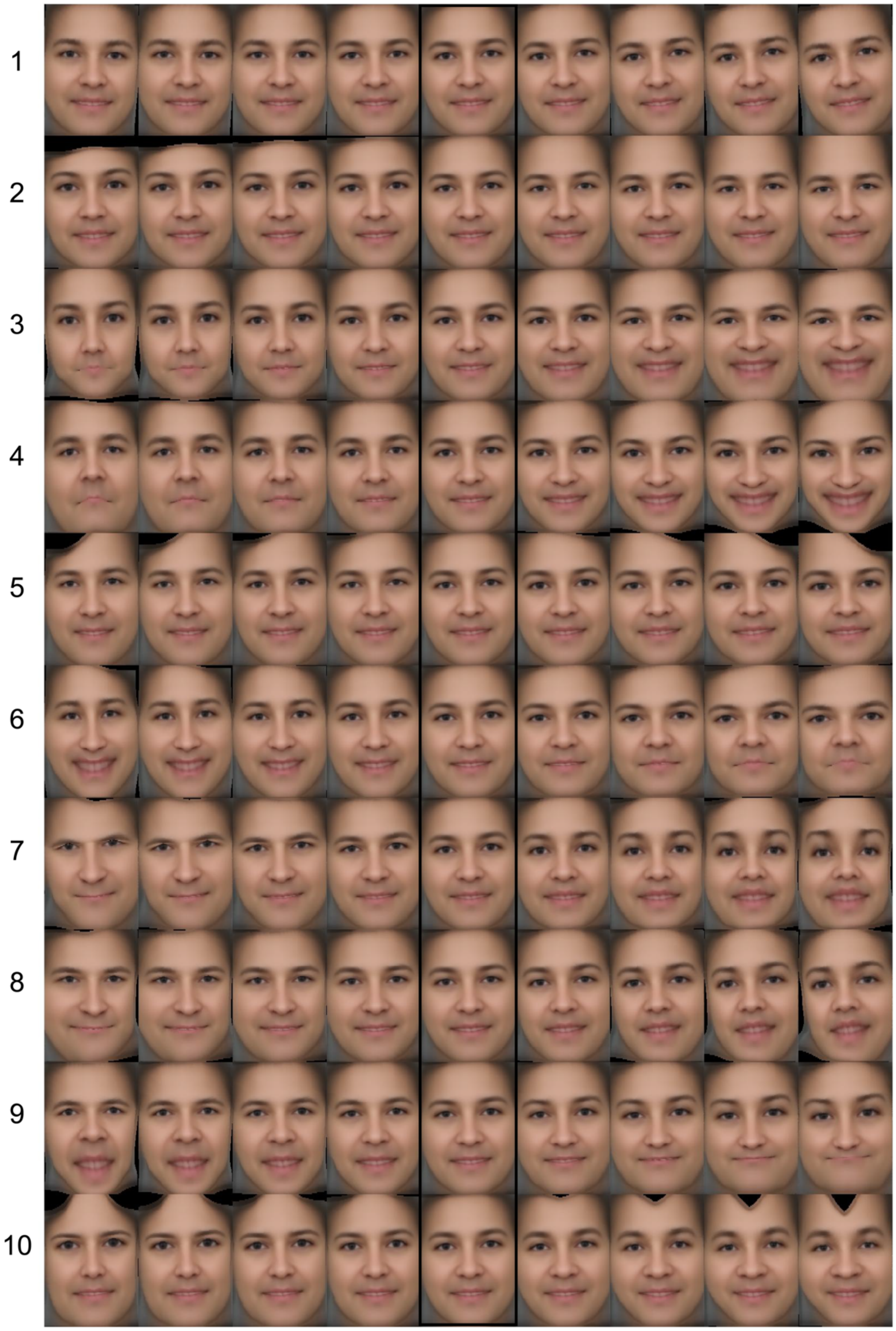

Figure 2.3. Illustration of the top ten shape components manipulated individually. The center column (black rectangle) is the mean face. Each row shows the mean face manipulated up (right) or down (left) on individual shape components. The components are ordered from highest (1) to lowest (10) variance explained. Collectively these ten components explain 86% of variance in landmark positions.

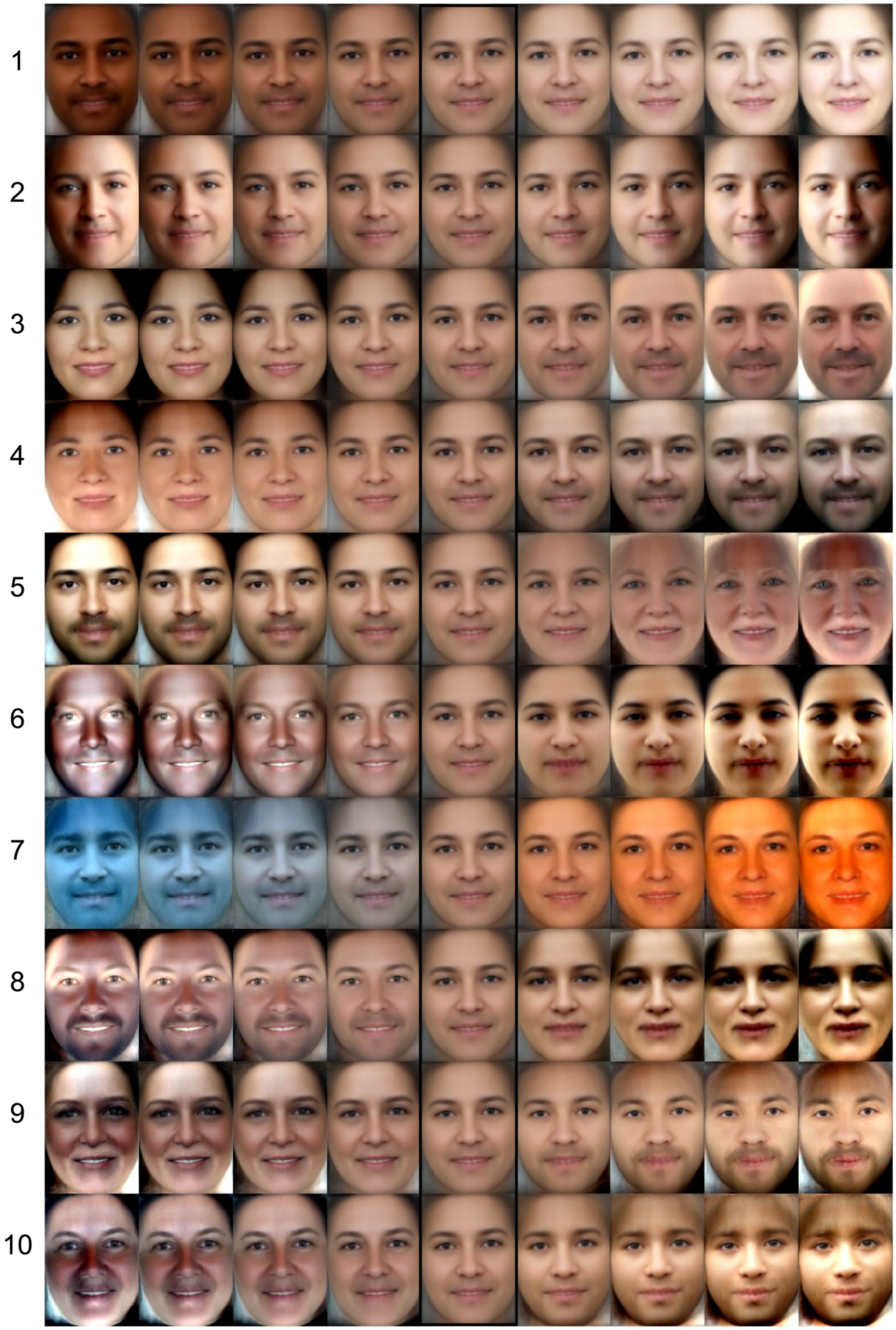

Figure 2.4. Illustration of the top ten appearance components manipulated individually. The center column (black rectangle) is the mean face. Each row shows the mean face manipulated up (right) or down (left) on individual appearance components. The components are ordered from highest (1) to lowest (10) variance explained. Collectively these ten components explain 69% of shape-free visual variance.

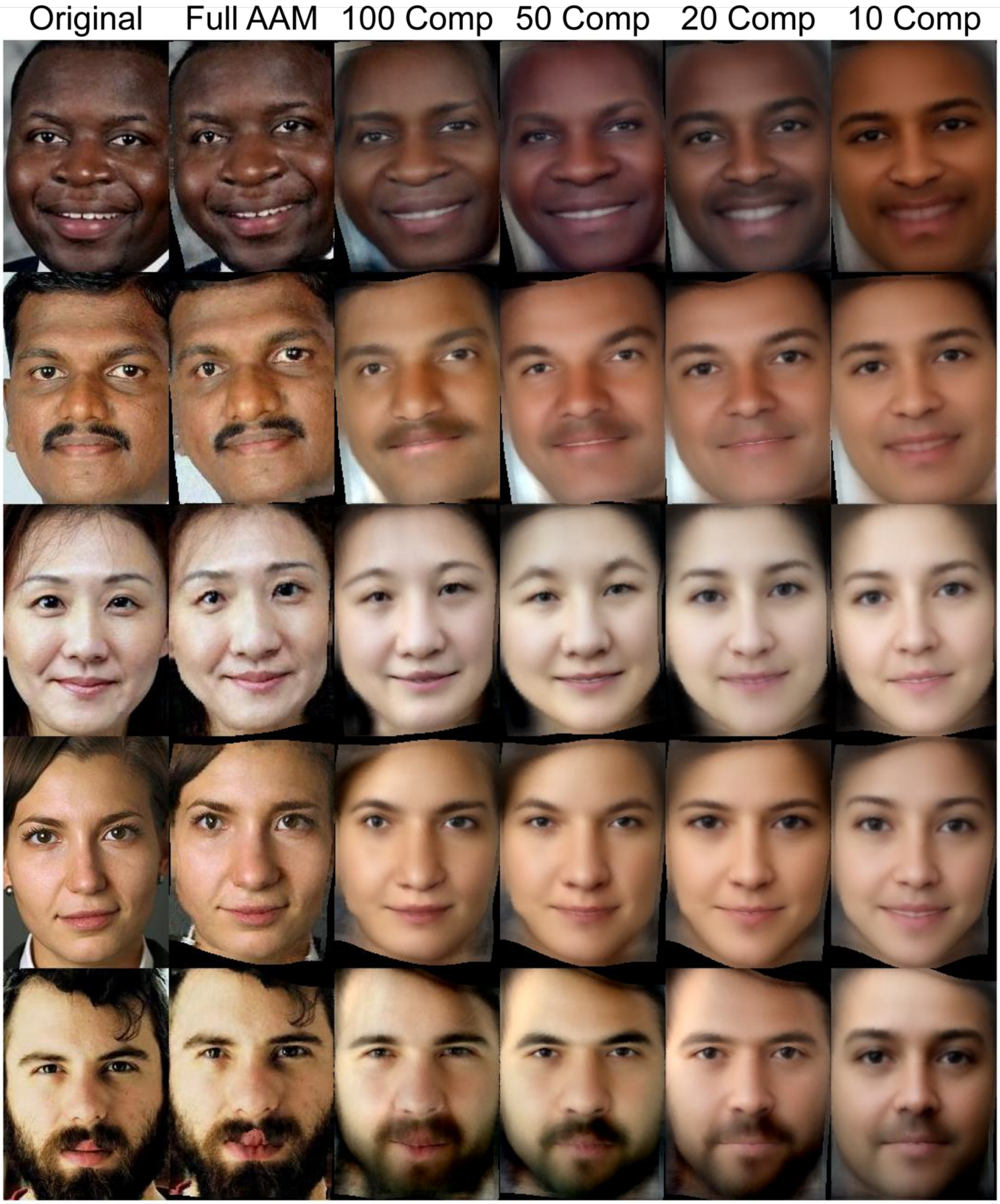

Figure 2.5. Example of five stimuli reconstructed with differing amounts of AAM components. The left column shows the original image. To the right of that shows the image reconstructed with differing amounts of AAM components included. First, "full AAM" was reconstructed with all 61 shape components and all 753 appearance components. The next four columns show reconstructions with 100 to 10 components included, with half coming from each type. For example, the 100 components column includes 50 shape components and 50 appearance components.

Subjective similarity sorting

The AAM allows for the creation of artificially generated faces based on changing or manipulating the value of the components. These components can be utilized to create stimuli with a controlled amount of similarity on one or more components. Furthermore, it allows for the active manipulation of stimuli during an experiment, either as controlled by the experimenter or interactively with the participant. However, due to the data-driven nature of the AAM, the components do not necessarily directly map onto perceptually-important dimensions. In order to make that mapping, we collected data on the perception of the faces. First, we analyzed data on the clustering of faces based on subjective similarity sorting.

A group of 6 research assistants acted as sorters of the face stimuli. The number of trials each sorter completed ranged from 87-520 (*M* = 218.67 ± 158.08, *Median* = 171.5). Each stimulus was presented at least once by 5 of the 6 sorters (the 6th sorter saw 1148/1156 stimuli). On average each stimulus was presented 73.8 times (± 2.91) to each sorter. Across all sorters, the average number of groupings created on each trial was 6.9 (± 1.77) out of a maximum of 10.

Using the sorting data collapsed across all sorters, we generated a dissimilarity matrix across all stimuli with each calculated as 1 minus the percentage of times each pair of stimuli were grouped together when they appeared in the same trial (see Fig. 2.6 for a visualization). In order to assess the reliability across all sorters, we first generated dissimilarity matrices based on every combination of 3 sorters (a total of 20 dissimilarity matrices). For each dissimilarity matrix in this set, there was one corresponding matrix with a set of 3 different sorters. As a measure of reliability, we calculated the correlation between all 10 non-overlapping pairs of matrices (Fig. 2.7). Any missing cells in either matrix pair were removed from the analysis. Of the 667,590 unique stimulus combinations in the full dissimilarity matrix, most were kept in this analysis (*M* = 481,317 ± 25,820, range: 438,311-512,307). On average the correlation between the dissimilarity matrices was 0.46 ± 0.05 (range: 0.37-0.52). This pattern indicates a consistent pattern of inter-rater reliability no matter the combination of sorters included. The unexplained variance could be reflective of differences between sorters, alternatively it could reflect how sorting is influenced by the unique combination of faces appearing on each trial.

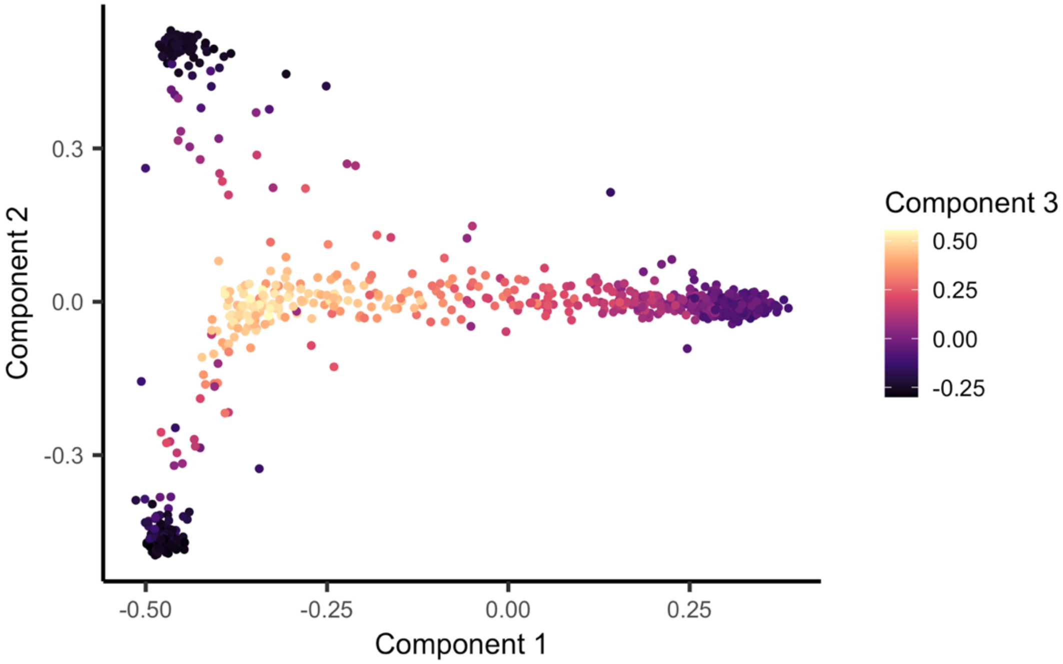

Figure 2.6. Top three multidimensional scaling (MDS) components across all face stimuli. The MDS analysis was run on the dissimilarity matrix generated from sorting the face stimuli.

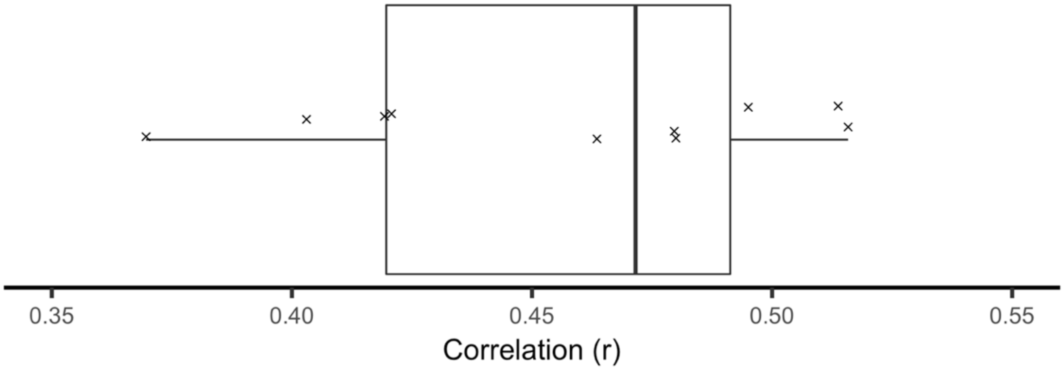

Figure 2.7. Boxplot of the correlation between dissimilarity matrices generated by every unique split of sorters. Individual correlations are indicated with an “x”. The correlation between the matrices were all near the mean of 0.46 ± 0.05.

As a demonstration of one application of this data, we proceeded to identify clusters of face stimuli that resembled one another most closely. We generated a distance matrix based on the Euclidean distance between the rows of the full dissimilarity matrix. We then performed a hierarchical clustering analysis on this distance matrix, using Ward's minimum variance method as implemented by the "hclust" function in R (Murtagh & Legendre, 2014). Based on a scree plot of the height when creating a different number of clusters, there were multiple logical cut-points depending on the intended purpose (Fig. 2.8,9). One way of visually inspecting the groupings is to look at the average face image across all stimuli included in a cluster. As one example, using 9 clusters (approach used in Chapter 3), the number of stimuli included in each cluster ranged from 58-293 (*M* = 128.44). Based on visual inspection, this cutoff point successfully generated distinct locations in face space (Fig. 2.10).

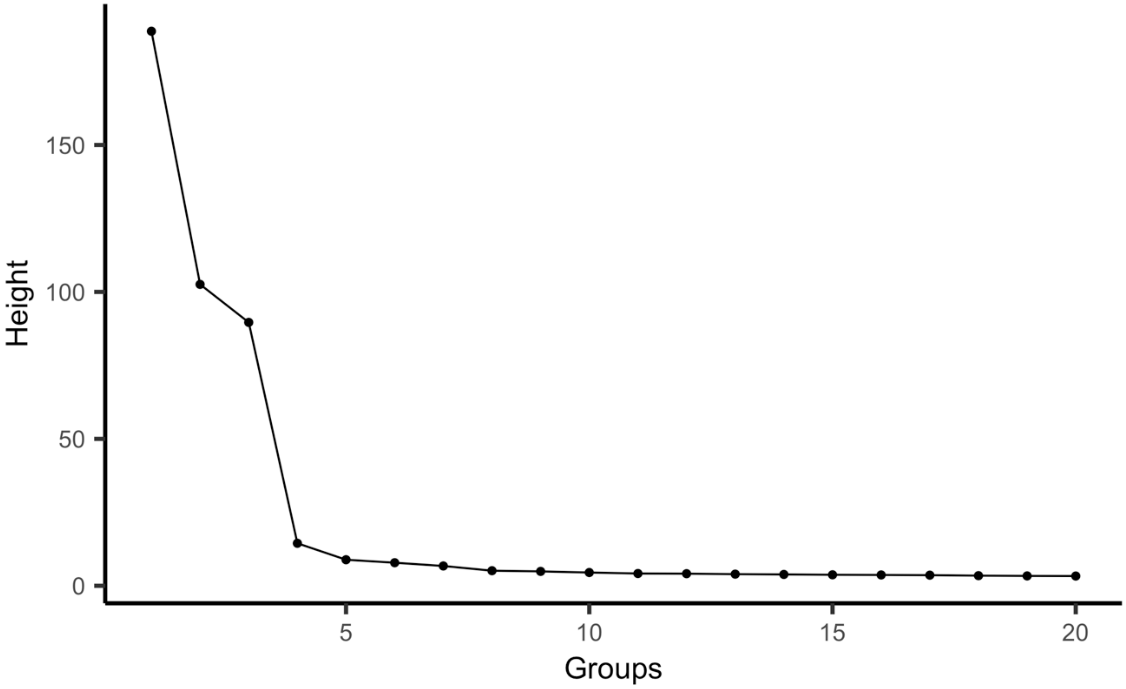

Figure 2.8. Scree plot of a hierarchical clustering analysis based on the distance between sorted face stimuli. 1-20 groups (x-axis) are included here, with the height of the groups plotted on the y-axis. Based on this plot, 4 is the most logical cut-point.

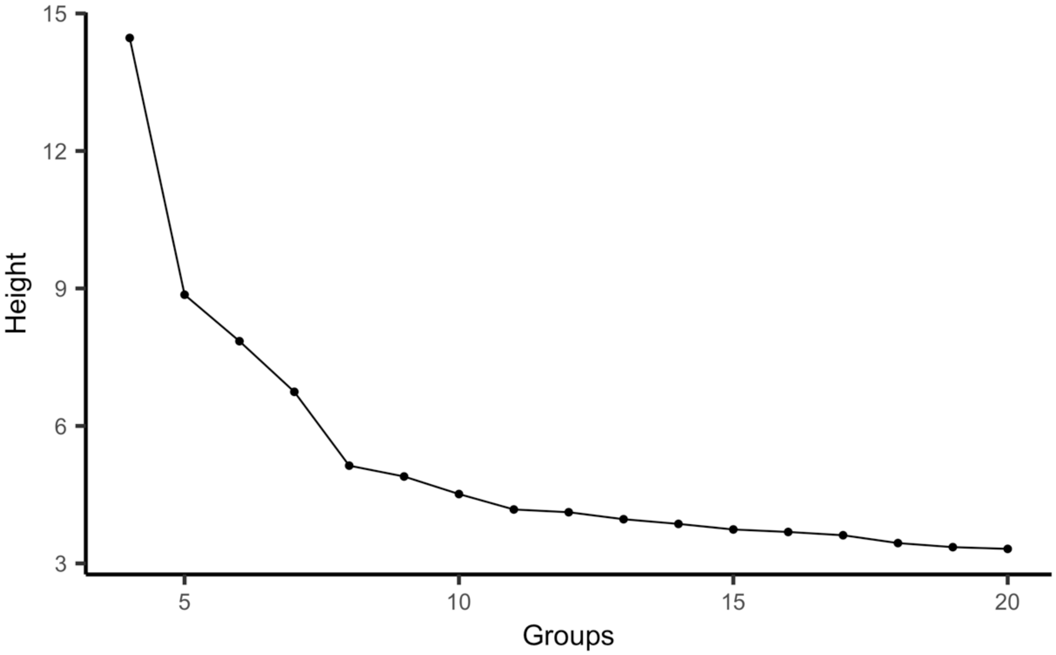

Figure 2.9. Scree plot of a hierarchical clustering analysis based on the distance between sorted face stimuli. 4-20 groups (x-axis) are included here, with the height of the groups plotted on the y-axis. By zooming in on a higher number of groups, additional logical cut-points emerge.

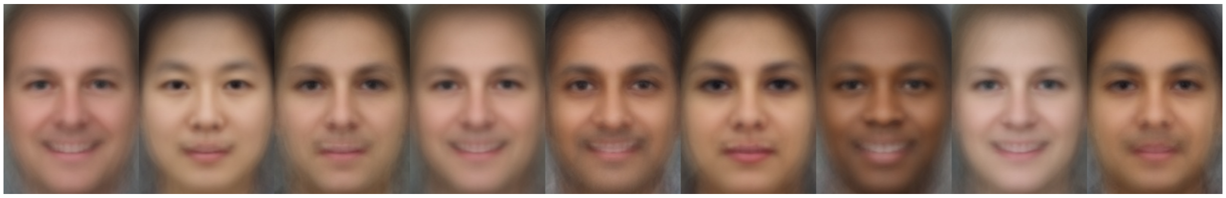

Figure 2.10. Example mean images of the face groupings, based on a hierarchical clustering analysis of the sorting data. In this example, the number of groups was set to 9.

Subjective ratings

Participants consistently utilized most or all of the range of the scale to make ratings (*M* = 8.31 ± 0.98; out of 9 maximum). Interestingly, there was some variation between dimensions with participants using slightly less of the range for attractiveness (*M* = 7.80 ± 1.01), dominance (*M* = 8.05 ± 1.23), and trustworthiness (*M* = 8.10 ± 1.02), but the full range for affect (*M* = 8.80 ± 0.523) and gender (*M* = 8.80 ± 0.523). This difference could be attributable to something about the stimulus set, or could be related to the more bivalent nature of affect and gender that could push responses away from the center of the scale. Participants not only responded at the extremes though, they tended to use every response within that range, with the number of distinct responses given closely corresponding to the range (*M* = 8.24 ± 1.07). The average standard deviation in responses was 1.87 ± 0.57.

Participants were consistent in the ratings they made, with high correlations between the ratings for repeated stimuli (dominance: *M* = 0.72 ± 0.13, range: 0.46-0.94; trustworthiness: *M* = 0.69 ± 0.15, range: 0.41-0.96; attractiveness: *M* = 0.78 ± 0.13, range: 0.52-0.95; affect: *M* = 0.88 ± 0.12, range: 0.67-0.98; gender: *M* = 0.88 ± 0.12*,* range: 0.51-0.97) and low average absolute differences (dominance: *M* = 0.78 ± 0.35, range: 0.10-1.45; trustworthiness: *M* = 0.85 ± 0.29, range: 0.41-1.52; attractiveness: *M* = 0.58 ± 0.20, range: 0.32-1.00; affect: *M* = 0.57 ± 0.25, range: 0.24-1.03; gender: *M* = 0.69 ± 0.22, range: 0.38-1.10).

In order to assess the inter-rater reliability of these ratings, we calculated the ICC on the consistency amongst raters for each rating and stimulus pool. The ICC was consistently high for affect (*M* = 0.74 ± 0.034, range: 0.72-0.79) and gender (*M* = 0.78 ± 0.039, range: 0.74-0.83). The other ratings were less consistently scored, as indicated by lower ICCs for dominance (*M* = 0.28 ±0.087, range: 0.18-0.36), trustworthiness (*M* = 0.38 ± 0.010, range: 0.28-0.50), and attractiveness (*M* = 0.40 ± 0.052, range: 0.35-0.46).

In order to prepare the ratings for future applications, we combined the ratings from each participant. First, ratings of stimuli repeated within participants were averaged together. The ratings were then z-scored within participants to help account for any differences in how participants utilized the rating scale. With the ratings now on a standardized scale, we averaged across participants.

The distributions and pairwise relationships between ratings is an important way to evaluate the validity of the data (Fig. 2.11). One initial validation is the bimodal distribution for gender (see Fig. S2.1). In fact, if classification were to be performed with the 0 point of the scale as the dividing line, classification accuracy was nearly perfect (female: 98.8%; male: 99.3%). Another important validation, is the high correlation (*r* = .72) between affect and trustworthiness. The low correlation between gender and affect (*r* = -.12) is a good validation that the face expressions in the set did not systematically vary by gender. One surprisingly strong relationship was between trustworthiness and dominance (*r* = -.56). Previous research has suggested that these are independent face dimensions (Oosterhof & Todorov, 2008). The relationship we found here could be driven by differences between the stimulus sets.

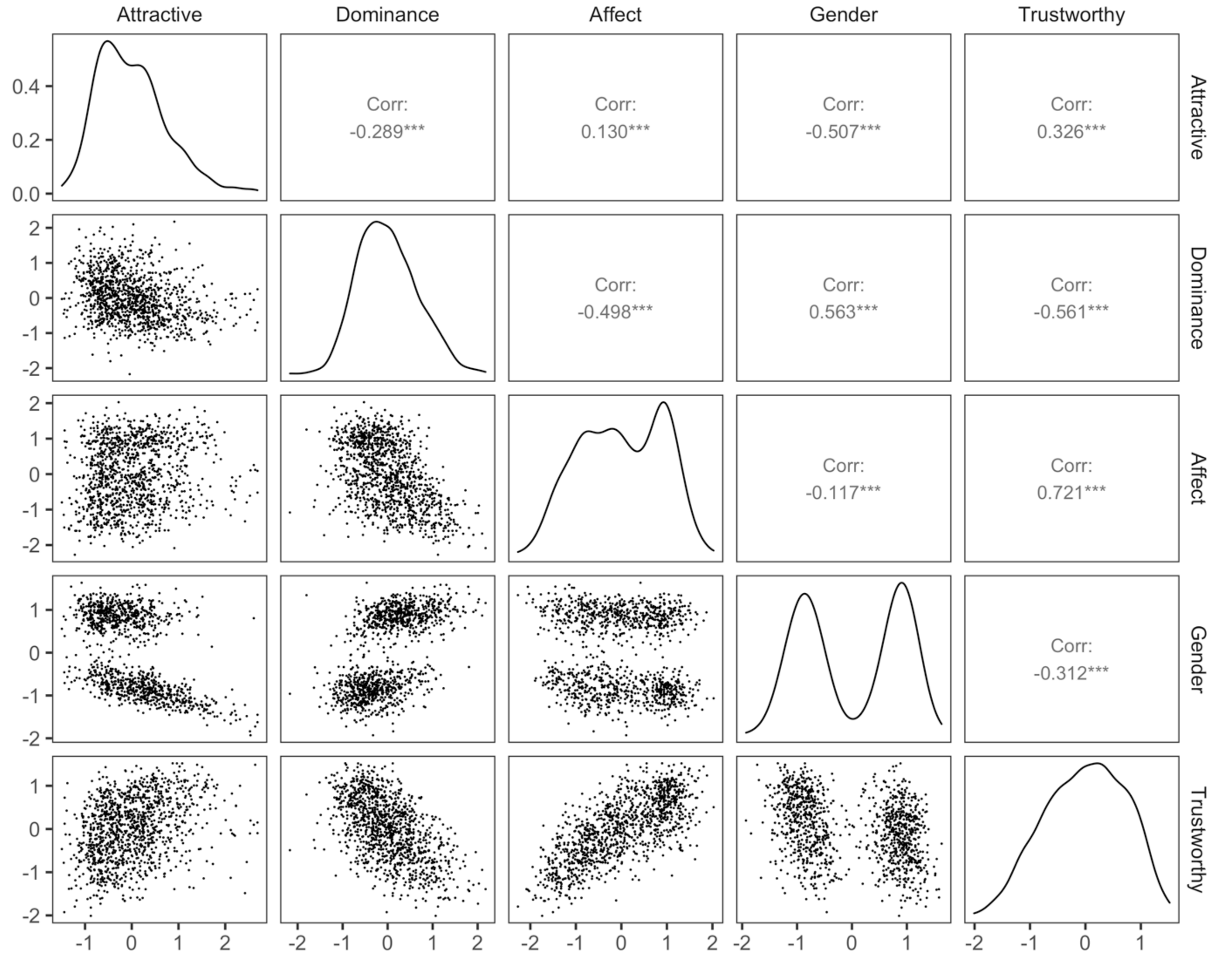

Figure 2.11. Relationship between the ratings for all stimuli. The ratings were z-scored within each participant and then average across participants. Top right: the pairwise correlations between each of the 5 ratings. Bottom left: the pairwise scatterplots between each of the ratings. Each dot represents one image. Diagonal: Density plot of each rating.

Discussion

Face images are a tremendously valuable stimulus class in psychological research. Here we make available a large corpus of face stimuli, all forward-facing, aligned, and uniform in size, but with a high degree of diversity on perceptually-important features. The stimuli have all been hand landmarked for use in AAM, have been sorted on similarity, and have been rated on perceptually-important feature dimensions. Combined, these attributes allow these stimuli to be utilized with a high degree of experimental control.

One of the main applications is the creation of synthetic face stimuli through the manipulation of the AAM components. With 61 shape and 753 appearance components, there is a large potential search space, with many locations in the space generating a combination of features that don't exist naturally. One way of choosing locations to generate realistic faces from is through mapping the grouping data to the AAM components (e.g. with a linear regression model). This allows for the generation of the "average" face from each group. In Chapter 3 we validated that faces generated from different groups were less like likely to cause interference with one another. This is useful for creating a stimulus design structure where interference only occurs where you intend it to.

The AAM components also allow for the experimental manipulation of specific components. When these components are combined with the subjective ratings, we can learn and utilize the relationship between the two. In one approach (see Chapter 3), we fit two regularized regression models with the AAM shape or appearance components as the outcome measures and all subjective measures collected as input variables. The weights from these models allow for the shifting of AAM components in relation to a specified shift in a subjective dimension (e.g. affect). This approach allowed for the creation of stimuli manipulated to be exactly the same, except slightly different on one perceptually-important dimension (see Fig. S2.2,3 for examples). Further, it allowed for the ability to probe memory on those same manipulated dimensions. For many experimental designs, this is an improvement over approaches such as warping, which rely on dimensions that need to be learned during the experiment. Other approaches, such as actors with different expressions, don't allow for a multi-dimensional and continuous search space.

There are times, however, when it may be preferable to employ natural face images. It remains unclear the extent to which the perception of synthetic face stimuli matches the perception of true face images. With recent advances in synthetic faces, there is evidence that they may be indistinguishable behaviorally (Shen et al., 2021), but that doesn't necessarily mean that they are the same in neural representational space (Dado et al., 2022). The present corpus has a diverse array of faces that is large enough to fill out the full range of the key perceptual scales included. Indeed, participants tended to use the full scale when making ratings. These ratings allow a quantification of differences between face stimuli on specific dimensions, allowing for experimental control without artificial manipulation. Further, the sorting data allows for the use of an overall similarity metric. Separately or combined, these metrics can be used in experiments that want to manipulate the similarity between stimuli or could be used in a category learning design based on the latent clustering of the face images.

The capability to have natural and synthetic images mapped onto the same latent dimensions further provides opportunities for integrated research approaches. For example, being able to map the synthetic and natural images into the same space allows for the possibility of testing whether there are differences in neural responses between them. Similarly, it allows for a more integrated research program where different experiments can use face stimuli from the same pool that can all be mapped into a shared feature space.

The stimuli and data we are currently making available are well equipped for research purposes, but there are areas to target for improvement going forward. Overall we found high reliability in our metrics, but there were some potential gaps. In particular, we found high reliability in most landmark positions both within and across raters. However, there were a small number of landmarks that had considerably lower reliability scores. We attribute that to the lack of variability in those particular landmarks, because features that the face stimuli were matched on (e.g. eye position) were the most likely to have low reliability. This lack of variation in those landmark positions, could lead to any errant deviation lowering the measured reliability. If that interpretation is true, we may either need a better metric to assess reliability or those landmarks are not adding any value to the AAM and should be left out.

Sorting reliability indicated high correspondence between different sorters, however there was some unaccounted for variance. The correlation between dissimilarity matrices derived from different combinations of sorters were all near the mean of 0.46. The unexplained variance could reflect differences between sorters, alternatively it could reflect how sorting is influenced by the unique combination of faces appearing on each trial. Evidence for either explanation could be found through the collection of additional sorting data. One barrier to the use of this data is that it was collected by research assistants rather than experimental participants. We plan to validate the current sorting results with independent experimental data.

Our subjective ratings demonstrated high inter-rater reliability for affect and gender, however the reliability was much more modest for trustworthiness, dominance, and attractiveness. It may be the case that these dimensions are more subjective and could vary by participant. The collection of additional data could help clarify whether this reflects a real difference between the ability to consistently rate these dimensions or is driven by the present participant pool. Collecting more data for these potentially less reliable dimensions could potentially make up for the variability in response patterns and reflect the average perception of that dimension.

There are a number of potential applications for this corpus of face stimuli. One important plan to further increase the utility of this face corpus is to make the fMRI data collected in this dissertation (see Chapter 4) publicly available. One example use case would be to generate a dissimilarity matrix based on the neural data and compare that to the behaviorally-derived one; this would provide a way to compare perceived face similarity between behavioral and neural-derived measures. Further, the neural dissimilarity matrix could be used as an overall similarity index for experimental design purposes as described above for the behavioral data. Each similarity metric could be useful in different experimental contexts, particularly since these similarity metrics are distinct both in terms of the measurement tool and the task. The neural data may ultimately act as a better pure reflection of overall similarity.

We make all stimuli, landmark positions, AAM components, sorting data, and subjective rating data available to other researchers. We hope that this provides a valuable set of face stimuli for a variety of experimental designs and purposes. We view them as particularly valuable in cases where multiple continuous face dimensions need to be quantified or manipulated. These attributes will help researchers utilize face stimuli to their full potential.

Chapter III

LONG-TERM MEMORY INTERFERENCE IS RESOLVED VIA REPULSION AND PRECISION ALONG DIAGNOSTIC MEMORY DIMENSIONS

From Drascher, M. L., & Kuhl, B. A. (2022). Long-term memory interference is resolved via repulsion and precision along diagnostic memory dimensions. *Psychonomic Bulletin & Review*, 1-15.

Introduction

When episodic memories are similar, this can lead to interference and forgetting. A critical point of emphasis in theories of episodic memory has been to not only characterize the contexts and situations in which interference occurs, but to consider the mechanisms that resolve interference (Anderson, et al., 1994; Anderson & Spellman, 1995; Anderson, 2003; Crowder, 2014; Fawcett & Hulbert, 2020; Smith & Hunt, 2000). To the extent that similarity is a root cause of interference, one potentially powerful way to reduce interference is to accentuate subtle differences between memories (Hulbert & Norman, 2015; Smith & Hunt, 2000). However, there is surprisingly little evidence characterizing whether or how the contents of episodic memories change as an adaptive response to interference.

One way to accentuate differences between similar memories is by increasing memory *precision*. For example, if two students look similar, more precise memories for the features of those students' faces (e.g., their specific eye colors) should render those memories more distinct. This concept is similar to the idea from perceptual learning that stimulus dimensions are 'stretched' to allow more fine-grained perceptual discriminations (Goldstone, 1998; Nosofsky, 1986). Analogously, increasing memory precision should expand the space between similar memories, thereby reducing interference.

An alternative, though not mutually exclusive, possibility is that differences between similar events are accentuated by *misremembering* event features as being more different that they actually were. For example, a pair of recent studies demonstrated that when otherwise identical objects were associated with slightly different colors, the color difference between those objects was systematically exaggerated in memory (Chanales et al., 2021; Zhao et al., 2021). Critically, this memory *repulsion* only emerged with extensive practice and coincided with reductions in interference-related memory errors. In fact, during early stages of learning, there was an 'attraction' in color memory (Chanales et al., 2021). Notably, repulsion-like biases have also been observed in working memory (Bae & Luck, 2017; Chunharas et al., 2018; Chunharas et al., 2019; Golomb, 2015) and visual attention (Chen et al., 2019; Won et al., 2020; Yu & Geng, 2019).

To the extent that episodic memory interference triggers changes in precision or bias, these changes should be most likely to occur (or most beneficial) along feature dimensions that are *diagnostic* of differences between similar memories. For example, if two students have identical hair color but slightly different eye color, then eye color would represent a diagnostic feature dimension. Targeted changes in discrimination accuracy along diagnostic feature dimensions have been observed during category learning (Goldstone & Steyvers, 2001; Kruschke, 1996; Theves et al., 2020) and in working memory (Chunharas et al., 2018). Computational models of episodic memory interference have proposed that episodic memory representations also undergo targeted changes that specifically exaggerate differences between similar memories (Hulbert & Norman, 2015), but empirical support for this proposal remains limited.

While precision and bias may both contribute to the resolution of memory interference, they are orthogonal constructs. Whereas precision refers to a reduction in memory variability, bias refers to a shift in a memory distribution. However, both measures require that memory be expressed using continuous values. Additionally, calculating precision requires that individual memories be sampled multiple times (to observe variability in the response). Despite recent progress towards utilizing continuous feature measures in episodic memory research (e.g. Berens et al., 2020; Brady et al., 2013; Cooper et al., 2019; Cooper & Ritchey, 2019; Harlow & Donaldson, 2013; Harlow & Yonelinas, 2016; Nilakantan et al., 2017; Nilakantan et al., 2018; Rhodes et al., 2020; Richter et al., 2016), prior studies have not specifically compared the relative contributions of precision and bias to the resolution of episodic memory interference.

Here, using multi-dimensional stimuli (faces), we tested whether similarity between stimuli induces adaptive changes in episodic feature memory (precision and/or bias) along diagnostic versus non-diagnostic feature dimensions. We developed a set of synthetic face stimuli that were manipulated on perceptually-important dimensions (Oosterhof & Todorov, 2008) as well as a behavioral face reconstruction task that allowed participants to express face memory by actively adjusting the synthetic faces. We used this innovative methodology across three experiments (including a preregistered third experiment) that each included a simple learning paradigm in which participants studied associations between faces and cue words (professions). Critically, most of the faces had a competitive *pairmate* that differed only on a counterbalanced diagnostic dimension (affect or gender). After extensive study and retrieval practice, we probed participants' memories for both feature dimensions simultaneously. Our central hypothesis was that competition would yield adaptive changes along the diagnostic feature dimension. Specifically, we predicted that memory for diagnostic features would be biased to exaggerate differences between similar memories (repulsion) and that repulsion would be associated with lower memory interference. We also predicted greater precision for diagnostic features and, importantly, tested whether repulsion and precision were independently predictive of memory interference.

Methods

We conducted three experiments with the same core experimental design and procedure. The only differences across the experiments were (1) the similarity of competitive pairmates increased very slightly from experiments 1 to 2 to 3, and (2) the minimum number of learning rounds was increased from experiment 1 to experiments 2 and 3 to account for the greater similarity/difficulty. Analyses and predictions for experiment 3 were preregistered (https://osf.io/s2gnq) after analyzing data from experiments 1 and 2. Thus, analyses are first reported for experiments 1 and 2, and then, separately, for experiment 3 (to test for replication). Exploratory analyses that combined data across experiments are also reported.

Participants

Participants were undergraduate students from the University of Oregon who received course credit for participation. A total of 40 participants were recruited for experiment 1. Four participants were excluded from analyses due to technical/procedural errors (see preregistration for full exclusion criteria: https://osf.io/s2gnq), resulting in a sample of 36 participants (M_{age} = 19.11 ±1.65, 18-25 years, 25 females). We sought a similar sample size in experiment 2 and recruited 41 participants (M_{age} = 20.49 ±2.47, 18-28 years, 28 females); no participants were excluded for technical/procedural errors. Based on the effect sizes in experiments 1 and 2 and corresponding power analyses, we recruited a sample 60 participants for a preregistered experiment 3 (see https://osf.io/s2gnq). Three participants were excluded for technical/procedural errors, resulting in a sample of 57 participants (M_{age} = 19.00 ±2.41, 18-22 years, 40 females). Each experiment involved a single session for each participant that lasted 90-120 minutes. Informed consent was obtained in accordance with procedures approved by the University of Oregon Institutional Review Board. All participants who were not excluded due to technical/procedural errors were included in our analyses of the associative memory test performance (see Procedure). Inclusion in all subsequent analyses was based on a set of performance-based exclusion criteria (see Performance-based exclusion criteria).

Materials

Cue words. For each participant and each experiment, the same set of 12 cue words was used (farmer, dentist, lawyer, teacher, chef, tailor, plumber, actor, artist, surgeon, judge, barber). Each cue word was assigned to a unique face, with the assignment randomized for each participant. All of the cue words referred to professions, consisted of one or two syllables, and were displayed in white with all capital letters.

Faces. Face images appeared in color with a uniform ellipse shape with a horizontal radius of 81 pixels and a vertical radius of 120 pixels. For all experiments, face images were generated from a set of eight *base faces*. The base faces were derived from a separate experimental procedure in which participants sorted a corpus of 1,008 faces into 'families' based on subjective assessment of the likelihood that faces were genetically related. Clustering algorithms were applied to the sorting responses to identify distinct clusters (families). Each of the eight base faces represents the mean face from a cluster, normalized for features not relevant to the grouping (see https://osf.io/6cew9/ for full details of stimulus generation methods). Critically, because of the way in which the eight base faces were generated, the base faces were distinct from each other according to characteristics that were orthogonal to the dimensions of affect and gender (which were the dimensions manipulated in the current experiments).

For each participant in each experiment, half of the base faces (four) were assigned to a *competitive condition* and half (four) were assigned to a *non-competitive condition*. The assignment of base faces to conditions was randomized for each participant. Base faces were manipulated along two dimensions—affect and gender—in order to generate the specific faces that participants studied (*studied faces*). For the four base faces assigned to the competitive condition, we created pairmates by generating two studied faces from each base face, with the common base being the source of competition. For the four faces assigned to the non-competitive condition, each base face was manipulated to generate a single studied face. Thus, a total of 12 studied faces were generated and used for each experiment.

For each experiment, each studied face was manipulated to fall into one of four locations in a two x two (affect x gender) space. That is, within each experiment, each studied face had one of two affect values and one of two gender values. To manipulate these dimensions, we collected subjective affect and gender ratings for all of the 1,008 faces in the corpus (see https://osf.io/znc58/) and then used regression analyses to learn the mapping between the gender and affect ratings and face image parameters (739 parameters in total) derived from an Active Appearance Model (AAM) (Chang & Tsao, 2017; Cootes et al., 2001; Edwards et al., 1998). Thus, the regression weights allowed for different affect and gender values to be translated to the 739-parameter feature space to manipulate the base faces. In order to maximize the independence of the affect and gender dimensions, for each of the AAM parameters, the dimension (affect or gender) with the highest magnitude regression weight was retained and the regression weight for the other dimension was set to 0. Thus, each face dimension (affect, gender) was associated with a distinct set of AAM parameters.

For the non-competitive condition, the four studied faces corresponded to the four locations in affect-gender space (one face per location), with the assignment of base faces to locations randomly determined for each participant. For the competitive condition, the eight studied faces again corresponded to the four locations in affect-gender space (two faces per location), with the assignment of base faces to locations randomly determined for each participant. Critically, the eight faces in the competitive condition included four sets of pairmates. For two of those sets, the pairmates within each set differed on affect and were matched on gender (i.e., diagnostic dimension = affect, non-diagnostic dimension = gender). For the other two sets, the pairmates differed on gender and were matched on affect (i.e., diagnostic dimension = gender, non-diagnostic dimension = affect) (see Fig. 3.1A). For the sets of pairmates that shared the same diagnostic dimension, each set corresponded to a different value on the non-diagnostic dimension, but the pairmates within each set had the same value on the non-diagnostic dimension. For example, for the two sets of pairmates for which gender was the diagnostic dimension, each set of pairmates would have a different value on the affect dimension, but the pairmates within each set would have the same value on the affect dimension.

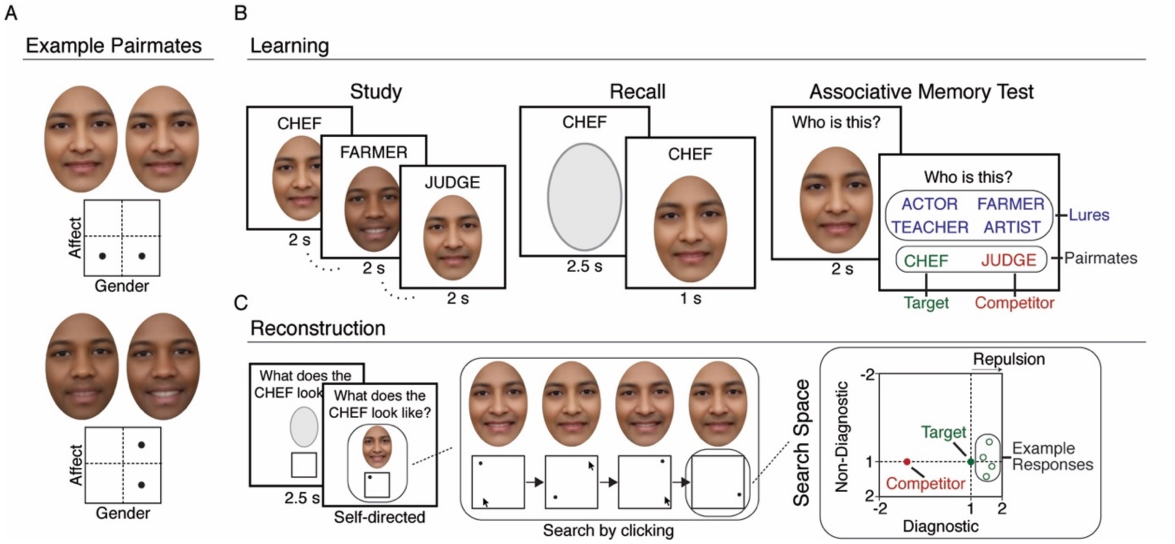

Figure 3.1. Experimental paradigm and design. **A.** Examples of competitive pairmates from experiment 1, with the location of the faces in affect-gender space shown below. Top: example of pairmates matched on affect (non-diagnostic dimension) but differing slightly on gender (diagnostic dimension). Bottom: example of pairmates matched on gender (non-diagnostic dimension) but differing slightly on affect (diagnostic dimension). **B.** Learning phase. Each round of the learning phase (up to 12 rounds total) consisted of three tasks. During study, participants viewed and studied associations between cue words and faces. During recall, participants viewed a cue word and were instructed to recall the corresponding face as vividly as possible; the correct face image then appeared. During the associative memory test, participants attempted to match each face image with its corresponding cue word, selected from a set of 6 options: target, competitor (the cue word of the pairmate face), and 4 lures (cues from other faces). **C.** Face reconstruction task. Left: participants were first shown a cue and instructed to visualize the corresponding face. Then, an altered version of that face appeared (shifted a random amount on the affect and gender dimensions). Center: participants used mouse clicks in a two-dimensional box to search the affect-gender space until the reconstructed face matched their memory for the target. Right: schematic of the search space showing the true location of the target (green dot) and competitor (red dot). Example reconstruction responses (open green dots) demonstrate our predictions: a bias away from the competitor (repulsion) on the diagnostic dimension and lower variability (greater precision) along the diagnostic compared to the non-diagnostic dimension.

For experiment 1, the difference between competitive faces along the diagnostic dimension was determined based on subjective assessment of the authors and initial pilot data. The goal was for the differences to be very subtle, yet learnable (see Fig. 3.1A for examples). Note: the units for these differences were not meaningful and are therefore not reported. For experiment 2, the difference between competitive pairs was reduced by 25% relative to experiment 1 in order to slightly increase the difficulty/interference. This was motivated by evidence that repulsion is more likely to occur when discrimination is relatively more difficult (Chanales et al., 2021). For experiment 3, the difference between competitive pairs on the gender dimension was the same as in experiment 2, but the difference on the affect dimension was reduced by 50% relative to experiment 1. This was motivated by evidence, from experiment 2, that interference was somewhat lower along the affect dimension compared to the gender dimension. Note that since the differences between competitive pairs in experiment 1 were quite small to begin with, the changes across experiments were subtle. For additional consideration of differences between affect versus gender across experiments, see Fig. S3.1.

Within each experiment, the difference between competing faces (pairmates) on the diagnostic dimension is described in *relative terms* (scaled units), with each face being 1 unit from the center of face space and, therefore, 2 units from each other. All faces were also exactly 1 unit away from the affect and gender borders in the response window (see Reconstruction phase, below). Analyses of face memory from the reconstruction phase were performed based on the distance, in units, between participants' responses and the actual location of the studied phases.

Procedure

Each experiment consisted of two main phases: a learning phase and a reconstruction phase. The purpose of the learning phase was for participants to extensively study and practice remembering the cue-face associations. The reconstruction phase served as the critical memory test for measuring bias and precision in face memory. All experiments were run in Matlab, using the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner et al, 2007; Pelli, 1997). All phases of the experiment had a gray background.

Learning phase. The learning phase consisted of up to 12 rounds, with each round split into two sub-rounds. Each sub-round included three blocks corresponding to the following experimental tasks, in the following order: study, recall, and associative memory test (Fig. 3.1B), with the exception that rounds one and two did not include the recall task. For each participant and each round of the learning phase, the 12 associations were randomly split into two groups of six associations each (four competitive, two non-competitive), with each group of six associations assigned to a separate sub-round. In other words, in each round of the learning phase, half of the associations went through study/recall/associative memory test and then the other half of the associations went through study/recall/associative memory test (with the exception, as noted above, that rounds one and two did not include the recall task). The rationale for splitting the associations into two sub-rounds was to facilitate learning by reducing the amount of information per block.

In the study task, participants viewed and studied the cue-face pairings. On each trial (2000 ms), a cue appeared directly above a face image. In between trials, there was a fixation cross for 200 ms. Participants were instructed to study the cue-face pairings; no response was made. In the recall task, participants attempted to recall the face associated with each cue. On each trial, a cue was presented above a blank ellipse (representing the to-be-recalled face) for 2500 ms. Participants were instructed to recall the associated face image as vividly as possible. Although no response was made, the correct face would then appear below the cue for 1000 ms as a way of providing feedback. In between trials, there was a 200 ms fixation cross. In the associative memory test, participants attempted to match face images with corresponding cue words. On each trial, a face image was presented for 2000 ms and was then replaced by a set of six different cue words displayed in the bottom half of the screen (three cues in each of two rows with the position randomly determined for each trial). The cue words included all of the cues from the current sub-round. For faces in the competitive condition, the set of cues included the correct answer (target), the cue that had been paired with the current face's pairmate (interference error), and four cues that had been paired with the other, unrelated faces (lures). For faces in the non-competitive condition, the set of cues included the correct answer (target) and five cues that had been paired with unrelated faces (lures). Participants made responses by clicking on the cue word with the mouse. After each response was registered, feedback indicated whether the response was correct ("Correct!"; 500 ms) or incorrect with the correct cue indicated (e.g. "Incorrect. This is the BARBER."; 2000 ms).

During the first two rounds of the learning phase, each study block presented each cue-face association three times. In subsequent rounds, each association was studied once per block. As noted above, there was no recall task in the first two rounds of the learning phase. In subsequent rounds, each association was recalled twice per recall block. Across all rounds of the learning phase, each association was tested three times per associative test block. For each task block (study/recall/associative test), the order in which each association was presented/tested was pseudo-randomly determined, with the following constraints: (1) all of the associations in each block were studied/presented once before any were repeated, (2) a given association was never presented/tested consecutively, (3) competing associations (face pairmates) were never presented/tested in consecutive trials. These constraints helped ensure that any comparisons between stimuli/associations were memory-based.

In experiment 1, participants repeated the learning phase for at least nine rounds and until they reached 100% accuracy on the associative memory test, up to a maximum of 12 rounds. Most participants had reached perfect accuracy after nine rounds (24/36), and nearly all did so after 10 rounds (31/36). Only two participants went through all 12 rounds, with one achieving perfect performance and the other being removed for continued poor performance (see below for performance-based exclusion criteria). In experiments 2 and 3, all participants completed 12 rounds of the learning phase regardless of associative memory test performance. For each experiment, participants were given the opportunity to take a break after every two rounds, with the length of the break determined by the participant. Participants were instructed to press the space bar when ready to proceed.

Reconstruction phase. After the learning phase, participants' memories for the features of the faces were probed with a surprise reconstruction task (Fig. 1C). On each trial in the reconstruction task, participants were first shown a cue (e.g. "What does the BARBER look like?") above a blank ellipse for 2500 ms and were instructed to bring the target face to mind. Next, an altered version of the target face appeared in the ellipse with a response box beneath the face representing the search space (see Reconstruction search space, below, for details). Participants used a mouse to click through the box; the face image above the box changed according to the location of each mouse click in the box. Although participants were not explicitly made aware of this, the box represented a two-dimensional affect-gender space. Participants were instructed to continue searching (clicking through the box) until the face matched their memory for the target face. Participants finalized their response by pressing the space bar. There was no limit on the response time. A fixation cross appeared for 200 ms between trials. Each of the 12 studied faces was probed (reconstructed) a total of four times in the reconstruction phase (48 trials total). The rationale for probing faces multiple times was so that the precision (variability) of reconstructions for each face could be measured. Faces were reconstructed in a pseudo-random block order. In each of four consecutive blocks (with no break or demarcation between blocks), each of the 12 faces was reconstructed once. As in the learning phase, the same face was never tested consecutively and pairmate faces were never tested in consecutive trials. After the reconstruction phase, there was a short phase where participants were prompted to provide a rating on a 9-point scale for both affect and gender for each stimulus. Results from this task (which was only included for validation) are not described here.

Reconstruction search space. In the reconstruction task, the altered face presented on each trial was derived from the same base face as the target face, but the affect and gender values were randomly selected from a range of possible values. This range of possible values corresponded to the size of the two-dimensional search space (i.e., the size of the response box). Importantly, the range of the search space and the center of the search space were identical across all trials, but the mapping of the dimensions to the x and y axes (e.g., x axis = affect, y axis = gender) and the direction/orientation of the axes (e.g., left = low, right = high) were randomly varied for each trial so that participants would not learn to associate a given face with a fixed spatial position in the response box. For each experiment, the size of the search space relative to the distance between pairmate faces was identical. That is, for each experiment the height and width of the search space was exactly twice the distance between pairmate faces on the diagnostic dimension. Thus, with pairmate faces 2 units apart (in our standardized units), the height and width of the search space was 4 units. For each trial, the location of the correct answer (target face) and the location of the pairmate face (for faces in the competitive condition) always corresponded to one of four possible locations (the center of each quadrant) with all four of those locations contained in the search space (see Fig. 1A).

Analysis methods

Performance-based exclusion criteria. For analyses that involved the reconstruction task data, we excluded a small number of participants based on performance during rounds 9-12 of the associative memory test. Participants were excluded if (a) their error rate for non-competitive trials was greater than 20% for any of these rounds or (b) they selected the lure faces on greater than 20% of the competitive trials for any of these rounds. Based on these criteria, one participant was excluded from analysis of the reconstruction task data in experiment 1 (yielding *N* = 35), four were excluded from experiment 2 (yielding *N* = 37), and eight were excluded from experiment 3 (yielding *N* = 49) (see https://osf.io/dj6q2/ for other exclusion criteria that were established but did not apply). The rationale for having a high threshold for inclusion of participants in the reconstruction task analysis was to minimize cases where participants reconstructed an entirely wrong face and to instead focus on bias/precision in otherwise correctly remembered faces.

Measuring associative memory. As noted above, the associative memory test was used to confirm that participants achieved high accuracy in associating cues with faces. The associative memory test also allowed for a manipulation check of whether the competitive condition induced interference (lower associative memory accuracy) compared to the non-competitive condition. Data from the associative memory test was first analyzed in terms of accuracy on competitive compared to non-competitive trials. We ran a separate repeated measures ANOVA for each experiment with factors of condition (competitive, non-competitive) and learning round (1-9 for experiment 1, 1-12 for experiments 2 and 3). For competitive trials, we also separated errors by whether they were attributable to competition (interference error) or not (lures). If errors were random, interference errors would occur on 1/5th (20%) of the error trials. To test whether interference errors occurred at above chance levels, we therefore ran one sample *t*-tests, for each experiment, comparing the mean percentage of interference errors (across all learning rounds) to 20%.

Measuring bias. As described above, on each trial in the reconstruction task the target face was located in one of four locations (the center of the four quadrants). Thus, for both the x and y axes of the search space, the target was half-way between the center and the border of the search space (Fig. 1A). To measure for potential bias, for each experiment all responses were aligned onto a common axis and rescaled onto a common scale, separately for each feature dimension (affect, gender). For the rescaled data, the range of possible responses for each dimension was -2 to 2, with 0 being the center of the face space (i.e., the center of the search space). For the competitive condition, the location of the target face on the diagnostic dimension = 1 and the location of the pairmate face = -1 (Fig. 3.1C). Thus, a bias *away* from the pairmate face would be represented by values greater than 1, whereas a bias *toward* the pairmate face (or toward the center of face space) would be represented by values lower than 1. For the non-diagnostic dimension, the location of the target face *and* the pairmate face = 1. Although faces from the non-competitive condition were included in the reconstruction task, bias was not measured for these faces because the distinction between diagnostic versus non-diagnostic dimensions did not exist. Rather, non-competitive faces were of critical importance in the associative memory test, where they served to establish an overall memory interference effect.

It is important to note that, for the reconstruction task, the response range on each trial was asymmetrically distributed around the target. If the response range had been symmetrically distributed around the target, then the correct response on each trial would have, by definition, been the center of the search space—which likely would have led participants to learn to simply respond in the center. However, the drawback of the approach we used is that, for the diagnostic dimension in the competitive condition, there was more opportunity to respond *toward* the pairmate face (values between -2 and 1) than *away from* the pairmate face (values of 1 to 2). Of course, this asymmetry works *against* our predicted effect of repulsion (values greater than 1). Nonetheless, in order to account for the asymmetrically restricted response range, we estimated the true mean by fitting truncated normal distributions to the data. For each participant, separate models were run for the diagnostic and non-diagnostic dimensions, with each model pooling data across faces and feature dimensions (affect, gender) in order to include a sufficient number of data points. Thus, each model included 32 data points (eight faces in the competitive condition x four reconstruction trials per face). Maximum-likelihood estimation was used to find the mean and standard deviation of a truncated normal distribution that best fit the data. The distributions were modelled using the truncnorm and MASS packages in R. We constrained the search space of the mean to a range of plausible values evenly balanced on either side of the target (+/- 1 unit) and constrained the standard deviation to be a maximum of 1 and a minimum of .1. Although we view the modelled means as a better estimate of the true means, there are some sources of variance that the models do not account for. For example, the models do not account for potentially unique distributions for each feature dimension and/or stimulus. Furthermore, there is evidence that there may be inherent, global biases in how face features are later recalled (Won et al., 2020; Bülthoff & Zhao, 2019). Critically, however, any global biases would equally influence the diagnostic and non-diagnostic dimensions. Therefore, our analysis primarily focused on *differences* in modeled means for the diagnostic versus non-diagnostic dimensions.

Measuring precision. In order to measure the precision with which diagnostic and non-diagnostic features were remembered *for each face*, we calculated the standard deviation of responses across the four reconstruction trials for each face, separately for the diagnostic and non-diagnostic feature dimensions. We then computed the mean of these standard deviation values for each participant, separately for the diagnostic and non-diagnostic dimensions.

Measuring the relationship between reconstruction bias and associative interference. In order to determine whether bias on the diagnostic feature dimension plays an adaptive role in reducing memory interference, we ran a series of mixed-effects models that focused on the relationship between bias measured during the reconstruction task and accuracy on the associative memory test (averaged across the last four rounds in order to capture the end state of learning). Although this analysis was performed at the level of individual items (faces), the accuracy value for each face was defined as the average accuracy for that face and its pairmate. As such, both pairmates with each set had the same accuracy value. The rationale for averaging accuracy across pairmates was that if, for example, participants associate two competing faces (pairmates) with the same cue word (profession), rather than treating one of these associations as ‘correct’ and the other as ‘incorrect,’ it is more appropriate for the error to be shared across the two faces.

For the analyses relating reconstruction bias to associative memory accuracy we excluded participants who had perfect accuracy, across all trials, on the final four rounds of the associative memory test. The rationale for this exclusion was that, for these participants, there was no variance in associative memory for the model to explain. Additionally, we did not run this analysis for experiment 1 given the near-ceiling performance on the associative memory test over the last four rounds (11 participants [31%] had 100% accuracy; and the remaining participants had mean accuracy of 95.96 ± 3.01% with an average SD within a participant of 3.62 ± 1.70). For experiments 2 and 3—which used more similar pairmates—associative memory accuracy was lower and, therefore, fewer participants were excluded due to ceiling performance (seven participants [19%] in e2 and six participants [12%] in e3; mean accuracy for the remaining participants, e2: *M* = 92.47 ± 7.58%, e3: *M* = 93.56 ± 6.26%).

For these models, it was critical to compute reconstruction bias at the level of individual faces. However, the method described above of estimating the average bias for each participant by pooling across trials/faces was not feasible for this analysis given the small number of observations (four trials per face). Thus, for this analysis we simply used the mean of the reconstruction response (across the four trials per face). In order to address the concern that any observed relationship between reconstruction bias and associative memory accuracy might be driven by potential ‘swap errors,’ our preregistered approach was to exclude any individual responses (trials) for which the scaled response was between -2 and 0 and to only retain responses for which the scaled response was between 0 and 2. For the diagnostic dimension, any responses that were closer to the competing pairmate than to the target were therefore excluded. All remaining responses were included in the mean response for each face. While rare, if a face was associated with an excluded response on all four reconstruction trials, that face was entirely excluded from analysis. For experiment 2, this occurred for a total of four faces distributed across four participants; for experiment 3, this occurred for a total of six faces distributed across six participants. While this preregistered approach for exclusion of potential swap errors was intended as a conservative approach for eliminating the influence of extreme errors, all of our main results remained significant when no responses were excluded. Additionally, in exploratory analyses that combined data across experiments 2 and 3, instead of excluding extreme responses altogether, responses between -2 and 0 were capped at a value of 0 which allowed for all trials to be retained in the model, but reduced the influence of extreme responses.

Mixed effects models were implemented in R using the lme4 package. Likelihood ratio tests were used to compare models with relevant variables to null models that excluded those variables. In order to account for potential differences related to whether the diagnostic dimension was affect versus gender, all models included this categorical variable as a fixed effect. In order to allow the relationship between reconstruction bias and associative memory accuracy to vary for each participant, we modeled the relationship between bias and associative memory accuracy with random intercepts and random slopes for each participant, where possible. Our preregistered approach to dealing with models that failed to converge or that reached a singular fit was to rerun the same model with the random slope for bias removed (see Barr et al., 2013). While all of our preregistered models did converge, an exploratory model which used the difference in bias on the diagnostic versus non-diagnostic dimension as a predictor failed to converge when a random slope was included; thus, we removed the random slope. Exploratory models that included only unsigned error or precision as predictors (without bias) failed to converge when random slopes were included for these variables; thus, we removed random slopes for these variables. Finally, exploratory models that included bias along with precision and unsigned error as predictors also failed to converge when random slopes were included for all variables; when removing random slopes, we prioritized retaining a random slope for bias, which led to the exclusion of random slopes for precision and unsigned error.

Results

Associative memory test

To test whether associative memory accuracy differed between the competitive and non-competitive conditions we conducted repeated measures ANOVAs for each experiment with factors of condition (competitive, non-competitive) and round (e1: the first nine rounds; e2 and e3: 12 rounds). For each experiment, there was a significant main effect of condition (e1: $F(1,35) = 26.14$, $p < 0.001$, $\eta_G^2 = 0.034$; e2: $F(1,40) = 67.43$, $p < 0.001$, $\eta_G^2 = 0.10$; e3: $F(1,56) = 88.21$, $p < 0.001$, $\eta_G^2 = 0.16$), with lower accuracy in the competitive condition (Fig. 3.2A). To confirm that this difference specifically reflected interference, we considered the types of errors made. For the competitive condition, errors could correspond to selecting the competitor face or one of the four non-competitive lures (Fig. 3.2B). If errors were random, the competitor would be selected on 1/5th of the error trials. However, combining error trials across rounds, the competitor was selected at above-chance levels (e1: $M = 60.18 \pm 19.68\%$, $t(35) = 12.25$, $p < 0.001$, $d = 2.04$; e2: $M = 71.29 \pm 15.78\%$, $t(40) = 20.82$, $p < 0.001$, $d = 3.25$; e3: $M = 78.63 \pm 11.58\%$, $t(56) = 38.21$, $p < 0.001$, $d = 5.06$), confirming that increased errors in the competitive condition reflected interference from the competitor face.

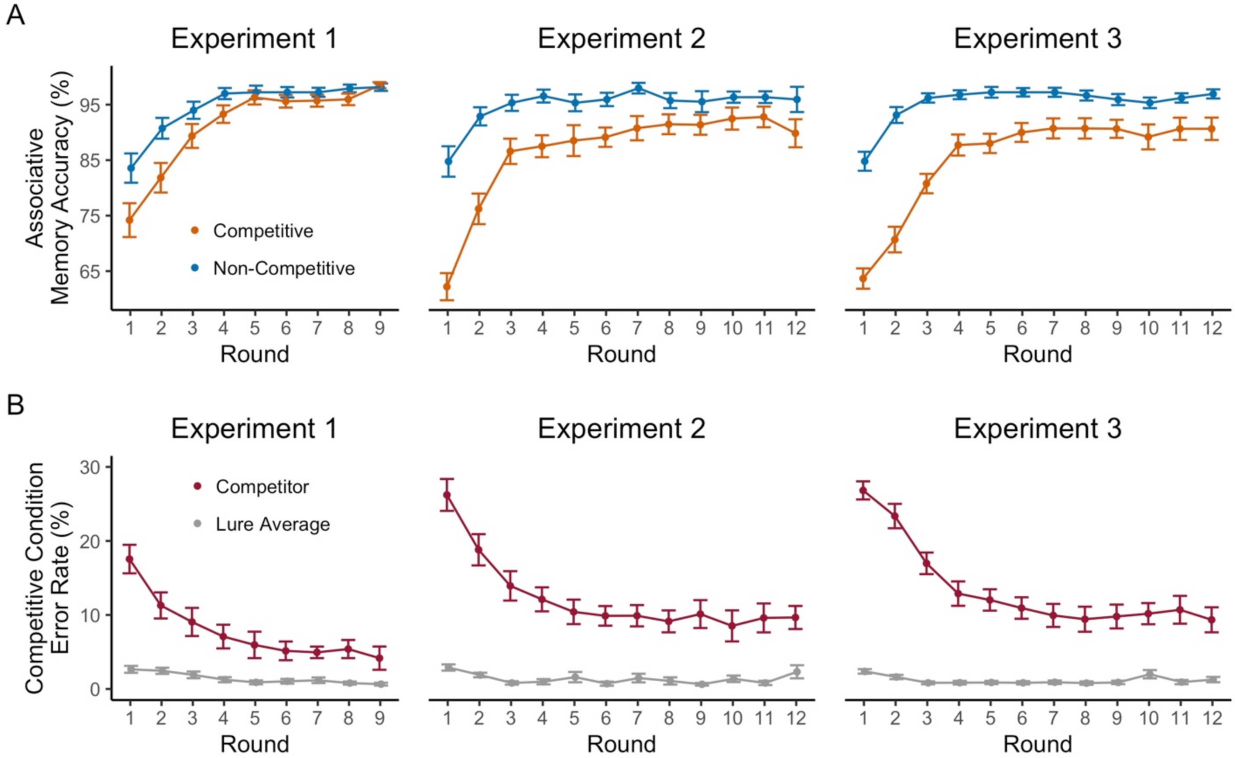

Figure 3.2. Associative memory test accuracy across learning rounds. **A.** Percent correct responses on the associative memory test during each round of the learning phase, separated by the non-competitive (blue) and competitive (orange) conditions and by experiment number. Performance was significantly higher for the non-competitive compared to the competitive condition in each of the three experiments. For accuracy in the competitive condition separated according to whether the diagnostic dimension was affect versus gender, see Fig. S3.1A. **B.** Error rates for the competitive condition on the associative memory test during each round of the learning phase. Data are separated by error type (competitor: red; lure average: grey) and experiment number. Competitors (the cues associated with the pairmate faces) were selected at a rate that exceeded the average rate of selecting one of four lures. Error bars represent SEM.

Face reconstruction accuracy

To test whether face reconstruction accuracy was above chance, we measured the Euclidean distance between each response and the target face location (in the two-dimensional response space; Fig. 3.1C). For each participant, the mean Euclidean distance between responses and target locations was compared against a permuted distribution (calculated by shuffling responses within participant 10,000 times). Above-chance accuracy (better than 97.5% of the permuted means) was observed for every participant (Fig. 3.3).

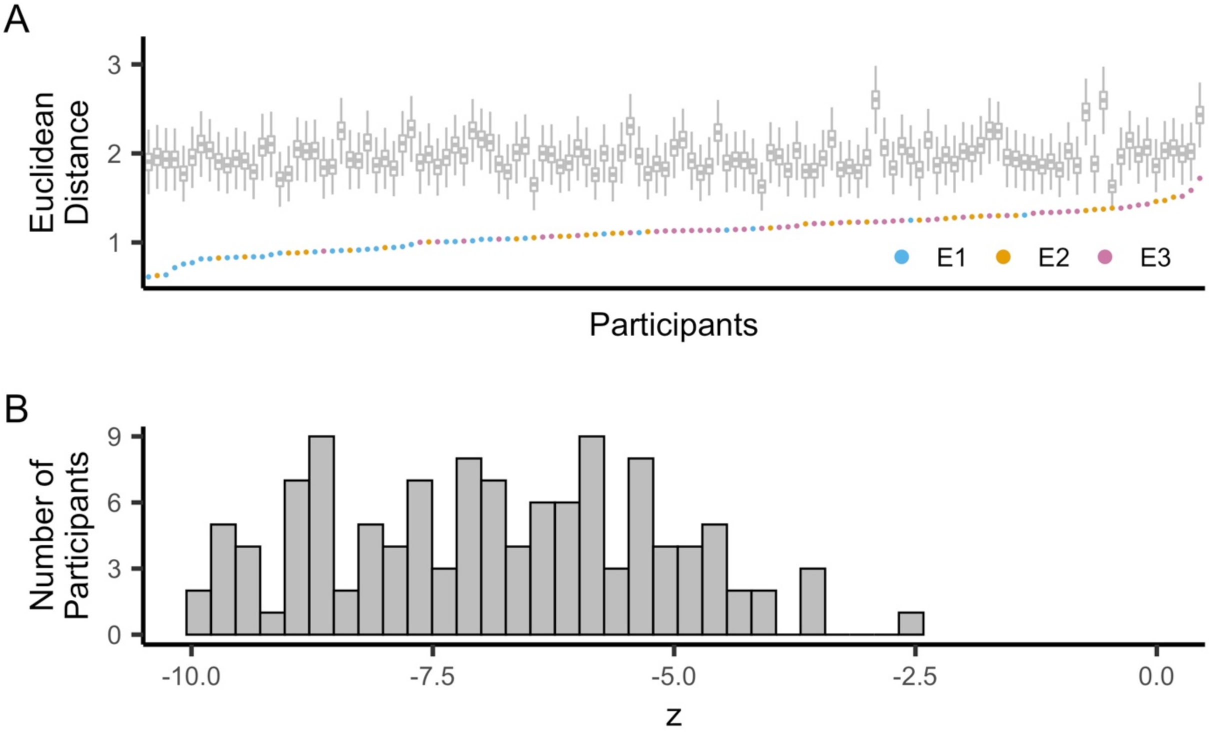

Figure 3.3. Face reconstruction accuracy. **A.** The mean Euclidean distance between the reconstructed location and the target location was significantly lower than chance for every participant as determined by comparing responses to a distribution of shuffled responses (10,000 shuffles per participant). The plot is arranged from participants with the lowest to highest mean Euclidean distance (left to right), with each participant represented by an individual dot (e1: blue; e2: orange; e3: pink). The distribution of shuffled responses for each participant is represented by a boxplot. **B.** Histogram of z scores reflecting each participant's mean Euclidean distance relative to the distribution of shuffled data (*M* = -6.87 ± 1.68, range: -9.97 - -2.59]). Lower z scores reflect better performance (lower Euclidean distance).

Face reconstruction bias

To test our critical prediction of repulsion along the diagnostic face dimension, we compared feature bias (see Methods) for the diagnostic vs. non-diagnostic dimensions in the competitive condition (Fig. 3.4A). We first tested predictions in experiments 1 and 2, and then tested for replication in experiment 3. A repeated measures ANOVA with factors of dimension (diagnostic, non-diagnostic) and experiment (e1, e2) revealed significantly greater bias toward repulsion on the diagnostic dimension ($F(1,70) = 22.25$, $p < 0.001$, $\eta_G^2 = 0.061$). There was a trend toward a significant interaction between dimension and experiment ($F(1,70) = 3.96$, $p = 0.0506$, $\eta_G^2 = 0.011$), with a relatively weaker effect size in experiment 1 ($d = 0.27$) than experiment 2 (*d* = 0.73). As predicted, experiment 3 replicated, with a large effect size and preregistered hypothesis, the greater bias toward repulsion on the diagnostic dimension (*t*(48) = 5.87, *p* < 0.001, *d* = 0.83).

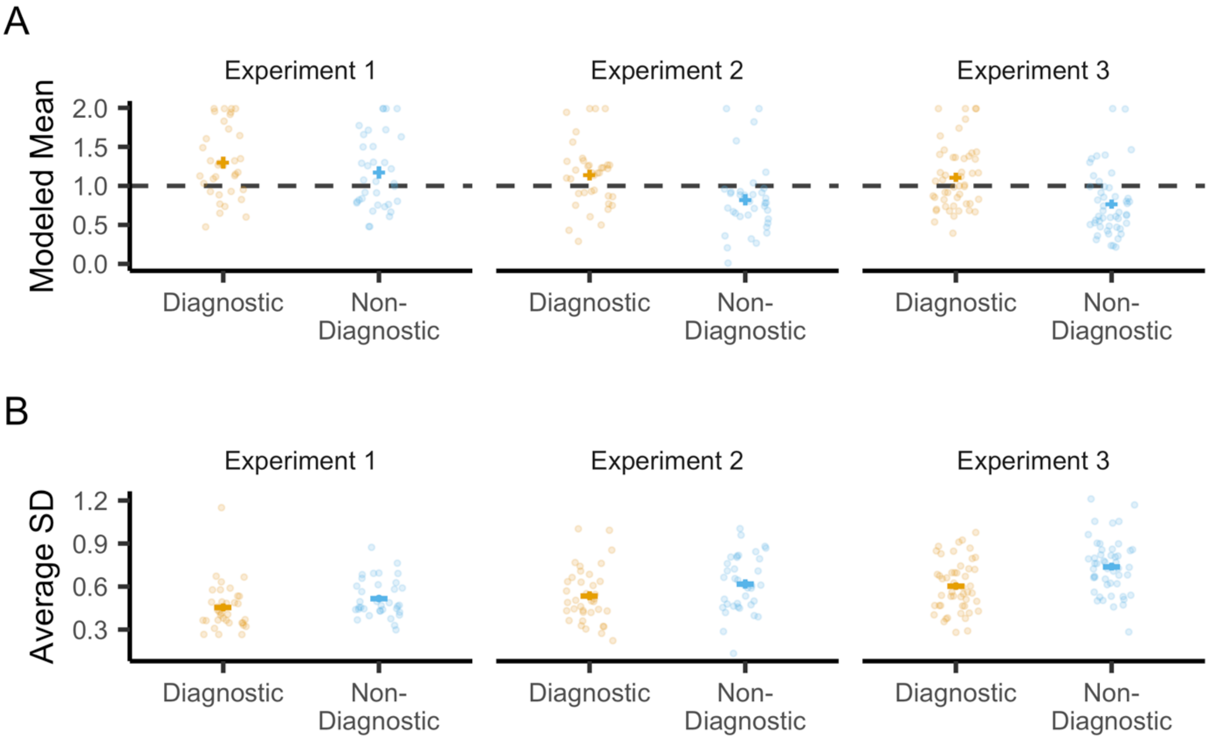

Figure 3.4. Feature memory from the reconstruction task along the diagnostic and non-diagnostic dimensions. **A.** There was greater bias towards repulsion (higher modeled mean response) on the diagnostic (orange) compared to the non-diagnostic (blue) dimension. **B.** There was greater precision (lower standard deviation of responses across the four reconstruction trials for each face) on the diagnostic compared to the non-diagnostic dimension. For analyses separated according to whether the diagnostic dimension was affect versus gender, see Fig. S3.1B,C. For analyses comparing the diagnostic and non-diagnostic dimensions with the non-competitive condition, see Fig. S3.3. Note: error bars represent SEM.

Although our preregistered analyses focused on the *comparison* between diagnostic and non-diagnostic dimensions, we also tested whether reconstructions on the diagnostic dimension significantly differed from the veridical location of target faces. Indeed, combining data across all three experiments, the modeled means for the diagnostic dimension were significantly greater than the true value of 1 (*t*(120) = 4.39, *p* < 0.001, *d* = 0.40), reflecting a bias away from the competing face. This effect did not significantly differ across experiments (*F*(2,118) = 2.15, *p* =

0.12, η_G^2 = 0.035). In contrast, on the non-diagnostic dimension there was a small, but significant bias toward the center of face space (modeled means < 1; $t(120) = -2.33$, $p = 0.021$, $d = 0.21$). This effect significantly differed across experiments ($F(2,118) = 9.56$, $p < 0.001$, η_G^2 = 0.14). In fact, in experiment 1 responses were significantly above 1 ($t(34) = 2.15$, $p = 0.039$, $d = 0.36$), and in experiments 2 and 3 they were significantly below 1 (e2: $t(36) = -2.45$, $p = 0.019$, d = 0.40; e3: $t(48) = -3.98$, $p < 0.001$, $d = 0.57$). While the absolute values of reconstructed responses should be interpreted with some caution (due to potential global biases), the consistent bias toward repulsion on the diagnostic dimension supports our prediction that competition triggers targeted repulsion on the diagnostic dimension.

Face reconstruction precision

We next tested whether reconstruction precision differed across diagnostic vs. non-diagnostic dimensions (Fig. 3.4B). We defined precision as the standard deviation across repeated reconstructions of the same face (see Methods). For the competitive condition, a repeated measures ANOVA with factors of dimension (diagnostic, non-diagnostic) and experiment (e1, e2) revealed significantly greater precision—i.e., lower reconstruction variability—on the diagnostic dimension ($F(1,70) = 16.81$, $p < 0.001$, η_G^2 = 0.044). This effect did not interact with experiment ($F(1,70) = 0.34$, $p = 0.56$, η_G^2 = 0.001). The effect of greater precision on the diagnostic dimension was replicated (consistent with our preregistered prediction) in experiment 3 ($t(48) = 5.45$, $p < 0.001$, $d = 0.74$).

Although our measure of precision was mathematically independent from our measure of bias, it is notable that these measures were correlated such that faces reconstructed with greater precision also tended to be associated with greater bias (see Fig. S3.2A). Importantly, however, the effect of greater precision on the diagnostic versus non-diagnostic dimension remained significant even when high-bias items were excluded from analysis (see Fig. S3.2B).

Relationship between reconstruction bias and associative interference

Finally, we tested our prediction that greater reconstruction bias (repulsion) on the diagnostic dimension is associated with better associative memory test performance (less interference). Due to near-ceiling associative memory performance in experiment 1 (Fig. 3.2), we focused on experiment 2 data. We ran a mixed-effects model that predicted item-level associative memory accuracy with fixed effects of (a) bias on the diagnostic dimension (continuous variable) and (b) whether the diagnostic dimension was affect or gender (categorical variable). Bias was modelled with random intercepts and slopes for each participant. Using a likelihood ratio test, we compared this model to a model without bias. Critically, model fit was significantly better when bias was included ($\chi^2(1) = 4.67$, $p = 0.031$), with bias positively predicting associative memory accuracy ($\beta_{bias} = 3.58$, $SE = 1.62$). As a control, we repeated the same analysis, but with bias on the non-diagnostic dimension; here, bias failed to improve model fit ($\chi^2(1) = 0.021$, $p = 0.89$, $\beta_{bias} = -0.31$, $SE = 2.14$). For experiment 3, we predicted (using a preregistered analysis) a replication of the relationship between diagnostic dimension bias and associative memory accuracy. We observed a small effect in the predicted direction, but it was not significant ($\chi^2(1) = 0.24$, $p = 0.63$, $\beta_{bias} = 0.69$, $SE = 1.41$).

In our preregistered analysis, we excluded reconstruction responses (trials) that were more similar to the competitor than the target. The rationale for this was to ensure that extreme responses (potential swap errors) did not have an outsized influence on the model (see Methods). However, this approach fully eliminated these trials rather than minimizing their influence. Therefore, as an exploratory analysis, we replaced these extreme reconstruction scores with a value of 0 (equal distance between the target and competitor, see Methods). This allowed all trials to be included, but reduced the influence of extreme responses (see Fig. S3.4 for further analysis of what these extreme responses may represent). For this exploratory analysis, we combined data from experiments 2 and 3, with experiment (e2, e3) added as a fixed effect. Compared to a null model, adding bias on the diagnostic dimension significantly improved model fit (χ^2(1) = 15.88, *p* < 0.001), with positive bias (repulsion) predicting higher associative memory accuracy (β_{bias} = 4.45, *SE* = 1.04). Adding an interaction between experiment and bias, did not improve model fit (χ^2(1) = 1.39, *p* = 0.24, $\beta_{expXbias}$ = -2.47, *SE* = 2.08), indicating that the relationship between bias and associative memory did not differ across experiments. Moreover, bias significantly improved model fit when applied to experiment 3 data alone (χ^2(1) = 3.98, *p* = 0.046, β_{bias} = 2.45, *SE* = 1.19), confirming that the relationship between bias and associative memory was not driven only by experiment 2 data. As a control, we ran the same model comparison but with bias on the non-diagnostic dimension as a predictor; there was no significant difference between models (χ^2(1) = 0.14, *p* = 0.71, β_{bias} = -0.40, *SE* = 1.08). Further, the degree of bias on the diagnostic dimension *relative to* the non-diagnostic dimension (i.e., the bias difference score) also significantly improved model fit compared to a null model without bias, χ^2(1) = 19.87, *p* < 0.001, $\beta_{bias.diff}$ = 2.71, *SE* = 0.60 (random slopes were excluded due to reaching singularity).

In an additional set of exploratory analyses that again combined data from experiments 2 and 3 we tested whether reconstruction bias on the diagnostic dimension predicted associative memory accuracy beyond what was predicted by unsigned error (absolute distance from the target on the diagnostic dimension) and precision (on the diagnostic dimension). Note: the following analyses did not include random slopes for unsigned error or precision (see Methods for rationale). Using hierarchical linear regressions with fixed effects of experiment (e2, e3) and feature dimension (whether the diagnostic dimension was affect or gender), model fit was significantly improved, compared to a null model, when unsigned error or precision were added (unsigned error: χ^2(1) = 16.42, *p* < 0.001, $\beta_{unsigned.error}$ = -5.72, *SE* = 1.40; precision:

$\chi^2(1) = 30.27$, $p < 0.001$, $\beta_{precision} = -5.91$, $SE = 1.06$). In other words, lower unsigned error and greater precision were associated with better associative memory. Critically, however, model fit significantly improved when bias was added to a model that already included unsigned error and precision ($\chi^2(1) = 4.39$, $p = 0.036$, $\beta_{bias} = 2.38$, $SE = 1.11$). Thus, bias predicted associative memory accuracy beyond what was explained by precision and unsigned error. Notably, model fit also significantly improved when precision was added to a model that already included unsigned error and bias ($\chi^2(1) = 26.51$, $p < 0.001$, $\beta_{prec} = -5.64$, $SE = 1.08$). Taken together, these exploratory analyses indicate that bias (repulsion) and precision—despite being correlated measures (Fig. S3.2A)—were independently predictive of associative memory performance (Fig. 3.5).

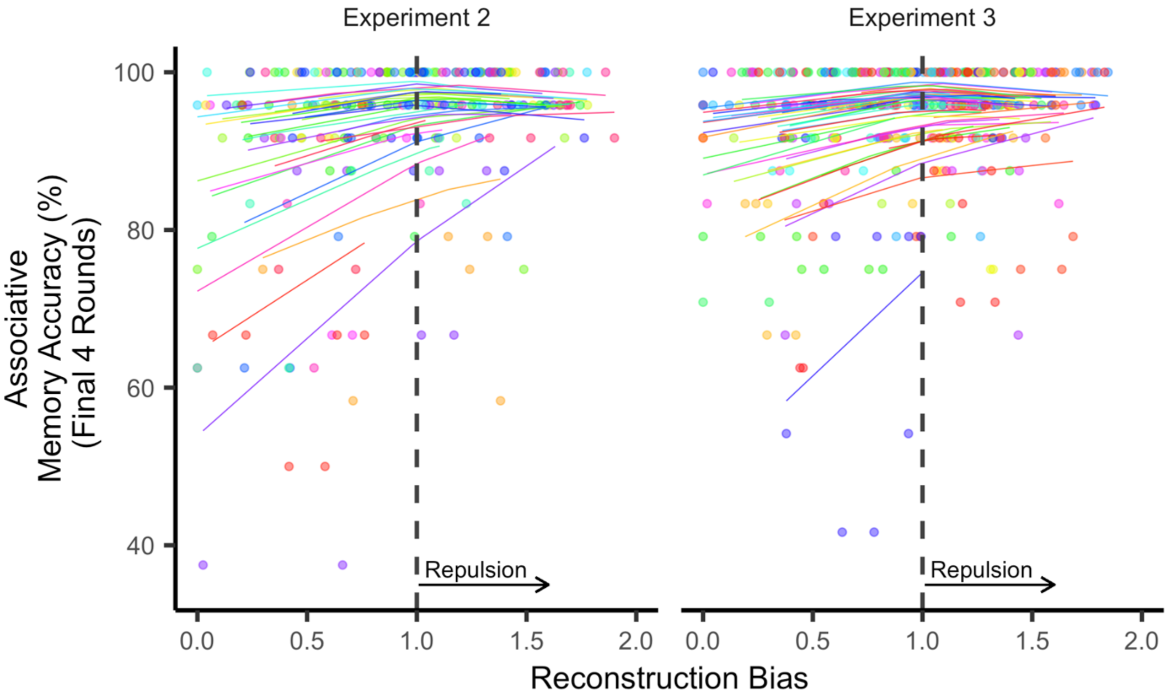

Figure 3.5. Relationship between reconstruction bias on the diagnostic dimension and associative memory accuracy. For the purpose of visualization, a mixed-effects model was run with mean associative memory accuracy (from the final four rounds of the learning phase) as the dependent variable and with experiment number, unsigned error, and bias included as predictors (gender/affect and precision were excluded). Stronger bias towards repulsion (reconstruction bias values > 1 reflect repulsion) was associated with higher associative memory accuracy (i.e., lower interference). Each dot represents a specific face image, with each participant plotted with a unique color. Each line represents the modelled, participant-specific relationship between reconstruction bias and associative memory accuracy. Note: bends in the lines reflect effects of absolute error.

Discussion

Across three experiments we found that similarity between long-term memories induced adaptive and feature-specific changes to the contents of those memories. We measured these changes using a two-dimensional face space (affect, gender), allowing us to separately measure memory along a dimension that was diagnostic of differences between similar faces and a dimension that was non-diagnostic of differences. We found that memory along diagnostic feature dimensions exhibited two key properties: (1) a systematic bias (repulsion) that exaggerated the difference between similar memories, and (2) greater precision (lower variability). Finally, we found that repulsion and precision were independently predictive of interference-related memory errors.

Although our paradigm was modeled after classic memory interference studies (Anderson, 2003; Anderson et al., 1994), the repulsion effect we observed is distinct from classic interference effects. If anything, interference predicts an *attraction* in remembered features. However, an important feature of our design is that face memory was only tested after extensive study and practice (Chanales et al., 2021, Zhao et al., 2021). Indeed, we found that greater repulsion in feature memory was associated with *lower* interference in the associative memory test. While it is important to note that this relationship failed to replicate using our preregistered analysis method in experiment 3, we view the updated method as a better approach for handling extreme responses, and the relationship we observed generalized across experiments and was independently significant in experiment 3. The relationship between repulsion and associative memory accuracy is notable when considering that repulsion fundamentally reflects a form of memory *error*. However, the error we observed was not randomly distributed; instead, it was systematically biased away from competing memories, thereby increasing the representational distance between memories. These findings complement evidence of conceptually-similar biases in working memory (Bae & Luck, 2017;

Chen et al., 2019; Chunharas et al., 2018; Chunharas et al., 2019; Golomb, 2015) and visual attention (Won et al., 2020; Yu & Geng, 2019). The ubiquity of these biases across domains suggests that repulsion is a fundamental, adaptive mechanism for resolving interference.

A central and novel focus of the present study was to compare repulsion along diagnostic versus non-diagnostic feature dimensions. The fact that repulsion was stronger for the diagnostic dimension provides important evidence that memories were not globally exaggerated (relative to the center of face space) in response to competition. Critically, in studies where only one featured dimension is probed (Chanales et al., 2021, Zhao et al., 2021), this interpretation cannot be ruled out. It is also noteworthy that because the mapping between affect and gender and the diagnostic and non-diagnostic dimensions was counterbalanced within participants, our results cannot be explained in terms of a bias along one feature dimension that generalized across all faces, as might occur in category learning (Goldstone, 1998; Goldstone & Steyvers, 2001). Finally, the relationship between repulsion and memory interference was selective to the diagnostic feature dimension, confirming that global biases were not adaptive. Thus, competition triggered targeted and adaptive distortions that preferentially occurred along the dimension that was essential for discrimination. These findings provide novel support for computational models of memory interference which propose targeted, feature-specific changes in memory representations (Hulbert and Norman, 2015; Norman et al., 2007; Norman, Newman, et al., 2006).

As with the repulsion effects, the precision effects we observed are in sharp contrast to typical interference effects. Specifically, whereas prior studies have shown that interference *reduces* precision in feature memory (Berens et al., 2020; Pertzov et al., 2017; Sun et al., 2017), our findings reveal that memory interference was associated with a relative *gain* in memory precision when comparing the diagnostic versus non-diagnostic dimensions. Importantly, however, we defined precision as the standard deviation across repeated tests of the *same* *memory*. This measure of precision was orthogonal to repulsion (or accuracy) as it reflected the consistency with which faces were remembered, regardless of the distance between remembered and actual values (absolute error). Put another way, if each face feature is represented by a distribution of potentially-remembered values, repulsion would reflect a *shift* in this distribution whereas precision would reflect reduced *variance* in this distribution (Yu & Geng, 2019). This is a key point because prior measures of memory precision have often assumed a distribution centered around the actual (veridical) memory value (e.g. Brady et al., 2013; Cooper & Ritchey, 2019; Harlow & Donaldson, 2013; Harlow & Yonelinas, 2016; Nilakantan et al., 2017; Nilakantan et al., 2018; Rhodes et al., 2020; Richter et al., 2016). While this is a reasonable assumption in many contexts, the current findings provide clear evidence, in the context of memory interference, that this assumption is violated.

An interesting avenue for future research will be to characterize the relationship between repulsion and precision. Here, these measures were mathematically distinct and were independently predictive of associative memory interference. Yet, repulsion and precision both have the consequence of increasing representational distance between competing memories and may therefore serve a common purpose. In fact, there was a robust correlation between these measures, with greater repulsion predicting greater precision (Fig. S3.2A). Thus, it is possible that repulsion and precision are distinct facets of a common underlying mechanism.

In summary, we demonstrate that episodic memories are modified and distorted in targeted and adaptive ways in response to interference. Whereas it is intuitive to conceptualize interference resolution as a reduction in memory errors, our findings support a distinct view in which systematic memory errors enhance discriminability between similar memories.

Chapter IV

RECONSTRUCTING FACE IMAGES
FROM DISTRIBUTED PATTERNS OF FMRI ACTIVITY
USING THE ACTIVE APPEARANCE MODEL

This chapter contains unpublished co-authored material. Maxwell L. Drascher is the primary author of this chapter with input from his advisor Brice A. Kuhl. Drascher and Kuhl designed the study together. Drascher supervised or conducted all data collection, and wrote the scripts for experiment presentation, data analysis, and figure creation. Data analysis was conducted with assistance from Paul Keene. Drascher wrote the manuscript with editorial assistance from Kuhl.

Introduction

Neural decoding has become an increasingly popular and important tool for neuroscientists (Norman, Polyn, et al., 2006). This approach to studying neural activity patterns began with broad, dichotomous decisions between perceptual categories (Carlson et al., 2003; Cox & Savoy, 2003; Haxby et al., 2001), but expanded its utility with the ability to make continuous feature predictions. This was first pursued with a focus on single, simple features such as the orientation of visual gratings (Ester et al., 2015; Kamitani & Tong, 2005; Serences et al., 2009), motion direction (Kamitani & Tong, 2006), and color (Brouwer & Heeger, 2009). More recently, researchers have utilized a neural reconstruction approach where complex, multi-dimensional stimulus classes are decoded across a set of dimensions that can collectively represent the image. Based on these decoded feature values, you can generate a neural reconstruction of the stimuli. This type of visualization can provide a window into what information is prioritized internally given a particular prompt and how that may differ depending on both experimental conditions and individual differences (Nestor et al., 2020).

Neural reconstruction approaches have often been applied to natural images (Beliy et al., 2019; Kay et al., 2008; Miyawaki et al., 2008; Mozafari et al., 2020; Naselaris et al., 2009; Seeliger et al., 2018) and movies (Nishimoto et al., 2011; Wen et al., 2018), but one stimulus class that has been of particular interest has been faces (Cowen et al., 2014; Dado et al., 2022; Güçlütürk et al., 2017; Lee & Kuhl, 2016; Nemrodov et al., 2019; Nestor et al., 2016; VanRullen & Reddy, 2019). Faces are of particular interest because humans are experts at perceiving and remembering faces (Kanwisher, 2000). This makes faces an important stimulus class for cognitive scientists, both in terms of studying how faces are processed as well as a stimulus class where small differences can be perceived and remembered (Nestor et al., 2020). Improving the ability to accurately reconstruct faces from neural activity would both offer important insights into face processing and would open up many experimental design possibilities.

Here, we approached face reconstruction from fMRI data in a similar way to some previous attempts (Cowen et al., 2014; Lee & Kuhl, 2016). In those attempts, a principal component analysis (PCA) was run on a set of face stimuli in order to generate a set of dimensions that describe each face in the set (eigenfaces). Here, we compare that approach to an updated approach to parameterizing the face images, the active appearance model (AAM). This approach previously performed better than eigenfaces at reconstructing face images based on electrophysiological recordings from macaque monkeys (Chang & Tsao, 2017). This updated approach also allows for face images to be represented with greater fidelity, with fewer components (50 vs 300 in the cited usages). The higher fidelity raises the ceiling on the ability to accurately and vividly reconstruct faces. The efficiency in representation makes subsequent analyses more efficient and interpretable; each component is more likely to represent features important for face perception and less likely to be representing other visual properties in the images. This increases the chance of successfully mapping brain activity patterns to these components, and may increase the likelihood of results generalizing across participants. Further, the AAM also introduces broad groupings of components (shape, appearance), that could help parse differences in face representation between brain regions (Chang & Tsao, 2017).

In the present study, across 35 participants, over 1,000 distinct face stimuli were viewed while in a scanner. These face stimuli were diverse in terms of gender, age, facial expression, race, and ethnicity, and have a wide range of information describing each, including AAM components and subjective ratings (see Chapter 2). Our approach here focuses on using a regularized regression algorithm to map patterns of fMRI activity while participants were viewing a face stimulus to individual face components. We focus on using independent models based on activity from different brain regions. In particular, we included regions that are important for early perception, face-processing, and also higher-level processing and memory. We compared the model's performance at predicting individual face components and component types across these brain regions. This approach also allows for the generation of reconstructed face images based on different regions, which provides a "view" into the participants' internal representation. Reconstruction accuracy was measured in terms of the distance between the predicted and true component values.

Methods

Participants

A total of 40 (M_{age} = 22.68 ± 4.17, 18-34 years, 24 females) right-handed, native English speakers, with normal or corrected-to-normal vision, from the University of Oregon community participated in the experiment. Five participants were excluded from analysis, four for having a high degree of head movement (greater than 10 instances of framewise displacement above 0.5 mm on multiple functional runs), and one for having a high non-response rate (greater than 20%

on 6/9 runs). This resulted in a final sample of 35 participants (M_{age} = 22.60 ± 3.95, 18-31 years, 20 females). Informed consent was obtained in accordance with procedures approved by the University of Oregon Institutional Review Board.

Procedure

Participants completed nine fMRI runs of a repetition detection task, each a total of 56 trials and for 7 m and 38 s. On each trial participants viewed an image centrally presented on the screen for 2 s, followed by a 6 s fixation cross (Fig. 4.1). Participants were instructed to pay attention to the images, because although most images were presented only once, a small percentage would be repeated. Using a button box, participants indicated whether each image was “new” or “old”. Responses were included if they were made within 7 s of stimulus onset. The button used (index or middle finger) to indicate new/old was randomly assigned to each participant. Each scanning run included 4 scene images and 48 unique face images, with 4 of those face images being repeated. The repeated images were used as the test images for image reconstruction, in order to have a better estimate of neural response to those images. No images were repeated across runs. The trial order was pseudorandomized within each run, with the constraint that test faces did not appear consecutively and there were at least 3 trials between repetitions of the same test face. Two of the runs were the same for all participants (fixed runs), in terms of stimulus inclusion and order, however the position of the run in the series of nine runs was randomized. All seven other runs were randomized within participant. In total, each participant viewed 432 unique images, with 396 used for training and 36 held out for testing. The test images were the same for all participants. The training images not included in the fixed runs were assigned to each participant in a pseudorandomized manner, with half being male/female.

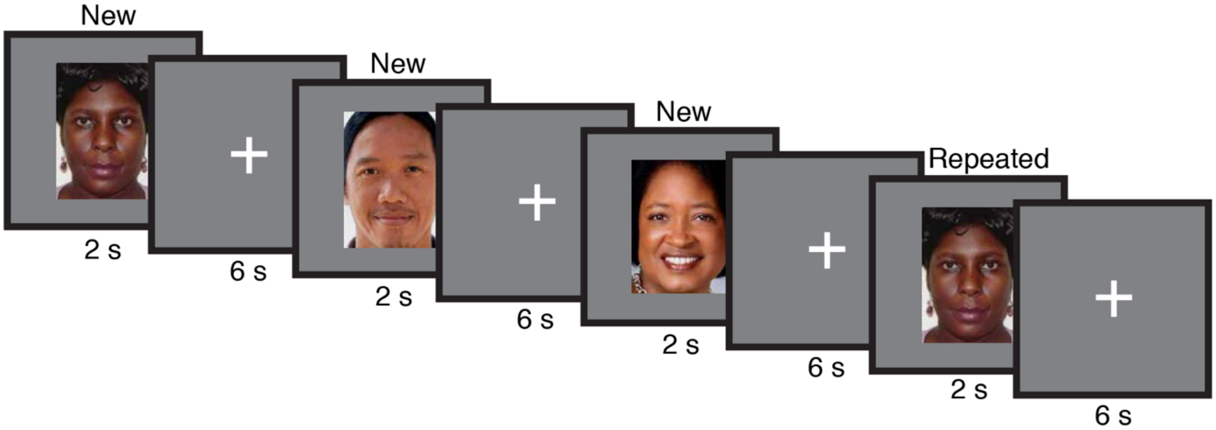

Figure 4.1. Experimental design. In the scanner, participants viewed images of faces or scenes one at a time and judged whether each image was "old" (repeated within the run) or "new" (novel).

An additional 10th functional run with the same task, but with a dynamic video of faces was included for many participants (31/40 total, 27/35 after exclusions), but that is not the focus of the present manuscript. Although the structure laid out above was the design, an error in the stimulus code led to some participants (14 total, 12 after exclusions) to accidentally be assigned stimuli from the fixed runs as training stimuli in other runs as well, which led to many faces being mistakenly repeated across runs for those participants (*M* = 65.58 ± 7.38, 54-82 stimuli).

Stimuli

A total of 1,148 faces were selected from a variety of online sources (see Lee & Kuhl, 2016). All faces were forward-facing and cropped and resized to 179 x 251 pixels. The faces were selected to be diverse in terms of age, race, ethnicity, and facial expression, with half being male/female. The full corpus is available at: https://osf.io/4uydh. A total of 36 images were pseudorandomly selected to be the test images based on including half male/female and including diversity in terms of race and facial expression. All other images were pseudorandomly selected to be included as training images for each participant, with half male/female and images not already used being prioritized. Of the 1,030 images in the training pool (outside of the fixed runs and testing), each stimulus was used at least once across all included participants (*M* = 10.0 ± 2.61, 1-18 times).

A small percentage of scene trials (four per run) were included to select face-preferring voxels. A total of 112 scene images were collected from freely available sources and cropped/resized to match the size of the face images. A diverse selection of indoor and outdoor scenes were included in the corpus. Scene images not included in the fixed runs were randomly assigned to each participant in a random order.

fMRI imaging acquisition

Imaging data were collected on a Siemens 3 T Skyra scanner at the Robert and Beverly Lewis Center for Neuroimaging at the University of Oregon. Whole-brain functional images were collected using a T2*-weighted multiband 229 EPI sequence (TR 2 s; TE 25 ms; flip angle 90°; grid size 104 x 104; voxel size 2 x 2 x 2 mm) and a 32-channel head coil. All participants had at least 9 functional scan runs, with most having 10 functional scan runs (see Procedure). After the functional runs, a whole-brain T1-weighted MPRAGE 3D anatomical volume (grid size 176 x 256 x 256; voxel size 1 x 1 x 1 mm) was also collected.

fMRI data preprocessing

fMRI data preprocessing was performed using fMRIPrep 21.00 (Esteban, et al., 2018), based on Nipype (Gorgolewski et al., 2011). A field map was estimated from two consecutive gradient-recalled echo acquisitions. The corresponding phase-map(s) were phase-unwrapped with prelude (FSL 6.0.5.1:57b01774). The T1-weighted (T1w) images were corrected for intensity non-uniformity with N4BiasFieldCorrection (Tustison et al., 2010) and skull-stripped using antsBrainExtraction.sh with OASIS30ANTs as the target template (ANTs 2.3.3; Avants et al., 2008). Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL; Zhang, Brady, & Smith, 2001). A T1w-reference map was computed after registration of 2 T1w images (after INUcorrection) using mri_robust_template (FreeSurfer 6.0.1; Reuter et al., 2010). Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1; Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (Klein et al., 2017). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the template (Fonov et al., 2009).

The estimated field map was then aligned with rigid-registration to the target EPI (echo-planar imaging) reference run. The field coefficients were mapped on to the reference EPI using the transform. Functional runs were slice-time corrected to half of slice acquisition range using 3dTshift from AFNI (Cox and Hyde, 1997). The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer; Greve & Fischl, 2009), using boundary-based registration with six degrees of freedom. Masks generated for each functional run were used to generate one mask based on the intersection of all masks. Functional data were smooth with a 2.0 mm FWHM Gaussian kernel using 3dBlurToFWHM from AFNI (Cox & Hyde, 1997). Each voxel was then standardized to a mean of 100 across time (within run), with values representing percentage signal change (in reference to the mean), and with a minimum of 0 and 200.

The data was modeled with a generalized least squares regression using AFNI's 3dREMLfit (Cox and Hyde, 1997). This analysis was performed by first generating a design matrix using AFNI's 3dDeconvolve function (Cox and Hyde, 1997). The design matrix included 6 motion parameters (x/y/z movement/rotation) as nuisance regressors, calculated using mcflirt (FSL 6.0.5.1:57b01774: Jenkinson et al., 2002). Framewise displacement (FD), was including as an additional nuisance variable, as calculated in Nipype (Power et al., 2014). Linear trends and low-frequency drifts were regressed out by including Legendre polynomials (4). Timepoints with a FD of above 0.5 were censored from this analysis. All nuisance regressors were regressed out at the run level. The hemodynamic response was modelled with a gamma function.

Regions of interest

All regions of interest (ROIs) were generated in a participant-specific manner based on FreeSurfer's Destrieux atlas (Destrieux et al., 2010; Fig. 4.2). These anatomical ROIs were co-registered to the functional images. We focused our analysis using three broad cortical ROIs, occipital (OCC), posterior parietal (PPC), and temporal (TEMP). We had no a-priori reason to expect meaningful or interpretable hemispheric differences, therefore all analyses used bilateral ROIs. OCC was defined as a combination of several regions across the occipital lobe that had little potential overlap with temporal or parietal regions (inferior occipital gyrus and sulcus, cuneus, middle occipital gyrus, superior occipital gyrus, occipital pole, calcarine sulcus, anterior and posterior transverse collateral sulcus, middle occipital sulcus and lunatus sulcus, superior occipital sulcus and transverse occipital sulcus, anterior occipital sulcus and preoccipital notch). The total number of voxels varied by participant (*M* = 6,467 ± 768, range: 5,204-7,979). PPC was defined as a combination of bilateral angular gyrus (ANG), intraparietal sulcus (IPS), supramarginal gyrus (SMG; combination of SMG and Jensen sulcus), and superior parietal cortex (SPC). The total number of voxels varied by participant (ANG: *M* = 1,830 ± 232, range: 1,463-2,357; IPS: *M* = 1,192 ± 145, range: 913-1,526; SMG: *M* = 1,832 ± 308, range: 1,375-2,640; ANG: *M* = 1,420 ± 221, range: 861-1,889). TEMP was defined as a combination of inferior temporal (Inf), superior temporal (Sup), middle temporal (Mid), temporal pole (Pole), and transverse temporal (Trans). Included with these other regions in some analyses, but not in the overall ROI, was the fusiform gyrus (FUS). The total number of voxels varied by participant (Inf: *M* = 2,103 ± 274, range: 1,400-2,525; Sup: *M* = 5,512 ± 586, range: 4,513-6,787; Mid: *M* =

2,149 ± 272, range: 1,523-2,742; Pole: *M* = 1,287 ± 183, range: 877-1,658; Trans: *M* = 126 ± 29, range: 45-190; FUS: *M* = 1,270 ± 173, range: 838-1,671).

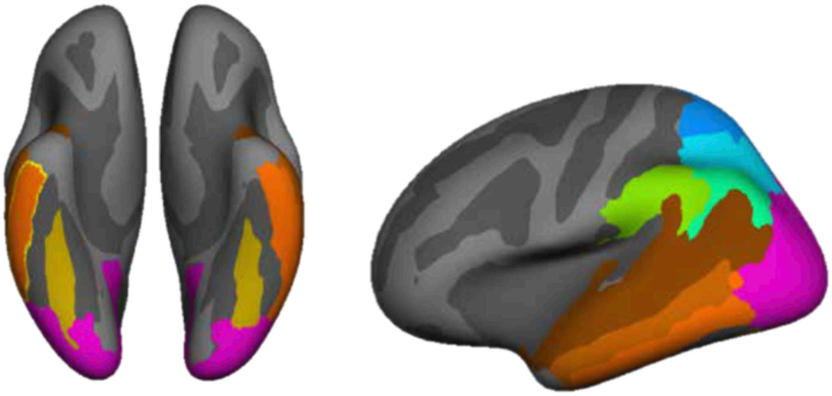

Figure 4.2. Visualization of the ROIs on the inflated surface of an averaged template brain supplied by FreeSurfer (magenta represents OCC; blue represents SPC; light blue represents IPS; green represents SMG; blue/green represents ANG; brown represents Sup; orange/brown represents Mid; orange represents Inf; yellow represents FUS). Left, ventral view. Right, lateral view.

Face reconstruction analysis

AAM components were generated based on the full corpus of face stimuli. Each face stimulus had 25 unique shape and 25 unique appearance component values. See Chapter 2 for full explanation of how these components were generated.

As a comparison face parameterization method, eigenface components were generated on the full stimulus set. Similar to AAM, eigenfaces are generated through a PCA analysis. Here, however, the PCA was performed on the raw image information (179 x 251 x 3 red/green/blue values), rather than performing a separate PCA on shape information first. Consistent with prior work, we focused on the top 300 components (Cowen et al., 2014; Lee & Kuhl, 2016).

For each participant and individual ROI, we performed a multinomial ridge regression using the beta values for all voxels within the ROI as predictors and the AAM components as the outcome values (Fig. 4.3). The regressions were performed using the “glmnet” package in R (Friedman et al., 2010). The weights relating the voxel activity and face components were then applied to the voxel activity evoked by the held out test images. Model predictions, unless otherwise noted, were made based on the beta values averaged across each trial including the same test image. For one set of analyses, we separated out the first compared to the second appearance of the test images.

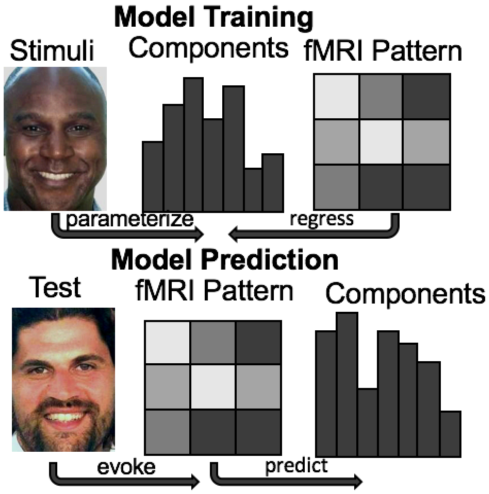

Figure 4.3. Schematic of the face reconstruction analysis. Top. The model was trained using a ridge regression that used the evoked fMRI pattern to predict the 50 AAM components for each face. Bottom. This model was then used to predict the AAM components based on the evoked fMRI activity patterns on a set of held-out test trials.

Face reconstructions were generated based on these predicted values (see Chapter 2 for reconstruction procedure). The accuracy of the reconstruction for each stimulus was determined by a series of two-alternative forced choice (AFC) tests. For each reconstruction, we measured whether it was more similar to the original face, or to a lure image. Similarity was measured by the Euclidian distance between the predicted and true component values. If the predicted component values were more similar to the true values than the lure values, the test was considered correct. The AFC accuracy for each image was based on the average of each

AFC test across using all test images as the lure image (35 total comparisons). The two face comparison makes chance performance 50%.

In order to assess performance at the individual component level, we used the same predictions as described above. For each participant, we then calculated the correlation between the predicted and true score across all 36 test images. We then converted correlation values into Fisher's z. Performance was assessed in reference to a correlation of 0.

For the subjective ratings (affect, gender, trustworthiness, dominance, and attractiveness), we ran a separate model, but with the same structure. Here, rather than predicting the individual AAM components, the five ratings were predicted. The subjective ratings are based on the average rating from MTurk participants (see Chapter 2). Performance on these predictions were evaluated in the same way as the individual AAM components above.

Statistical tests

Performance compared to chance was assessed using one sample t-tests for each separate model. Statistical tests were assessed at the 0.05 alpha threshold. No corrections have been made for multiple tests.

Results

Behavioral performance

Overall, participants performed well at the repetition detection task and were engaged. Participants began the task with high accuracy in block 1 (*M* = 89.7% ± 16.6%) and steadily improved through block 9 (*M* = 93%, ± 7.2%; overall: *M* = 91.5% ± 6.3%). The mean sensitivity (*d'*) was 2.60 ± 0.73. The average response time was 1765 ms ± 502 ms.

Reconstruction of faces using AAM

Reconstructions were generated based on models run separately for each ROI and participant (see Fig. 4.4 for example reconstructions). We focused our initial analyses on the three main overall ROIs (Fig. 4.5). In order to quantify the quality of the reconstruction, we computed the similarity between the predicted AAM components and the true components and compared this against the similarity between the predicted components and another test image's true components in a series of 2-alternative forced choice (AFC) tests. If the predicted components were more similar to the true components than the comparison image components, the test was considered correct. To assess performance compared to chance (50% accuracy), we ran a series of one sample t-tests for each ROI. All three ROIs were significantly above chance (OCC: $t(34) = 6.85$, $p < 0.001$, $d = 1.16$, $M = 53.4\%$; PPC: $t(34) = 3.59$, $p = 0.001$, $d = 0.61$, $M = 51.1\%$; TEMP: $t(34) = 3.48$, $p = 0.001$, $d = 0.59$, $M = 51.2\%$). The three regions, however, significantly differed in their level of accuracy ($F(2, 68) = 19.03$, $p < 0.001$, $\eta_G^2 = 0.17$), with OCC performing significantly better than PPC ($t(34) = 5.2$, $p < 0.001$, $d = 0.92$) and TEMP ($t(34) = 4.67$, $p < 0.001$, $d = 0.82$). Performance in PPC and TEMP did not significantly differ ($t(34) = -0.56$, $p = 0.58$, $d = 0.10$).

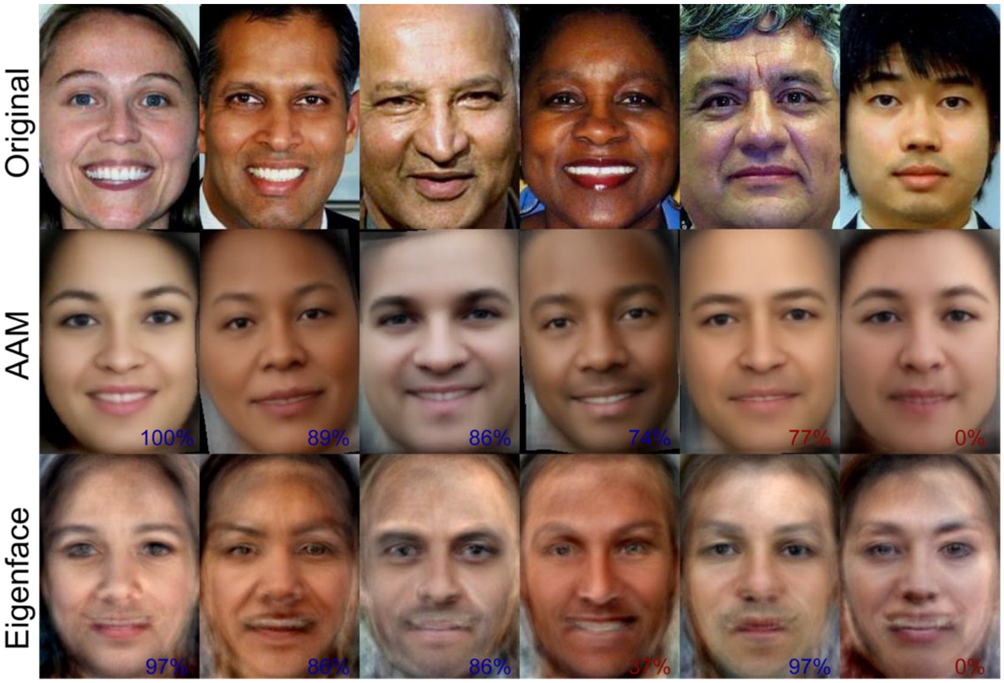

Figure 4.4. Reconstruction examples from one participant from OCC. The reconstructions were based on both AAM (middle) and eigenface (bottom) components compared to the true image(top). The AFC accuracy for each reconstruction is in the bottom right corner. The first three columns (left to right) are examples where both face models performed well, the next two are examples where the models diverged in their performance, the final column shows poor performance for both models.

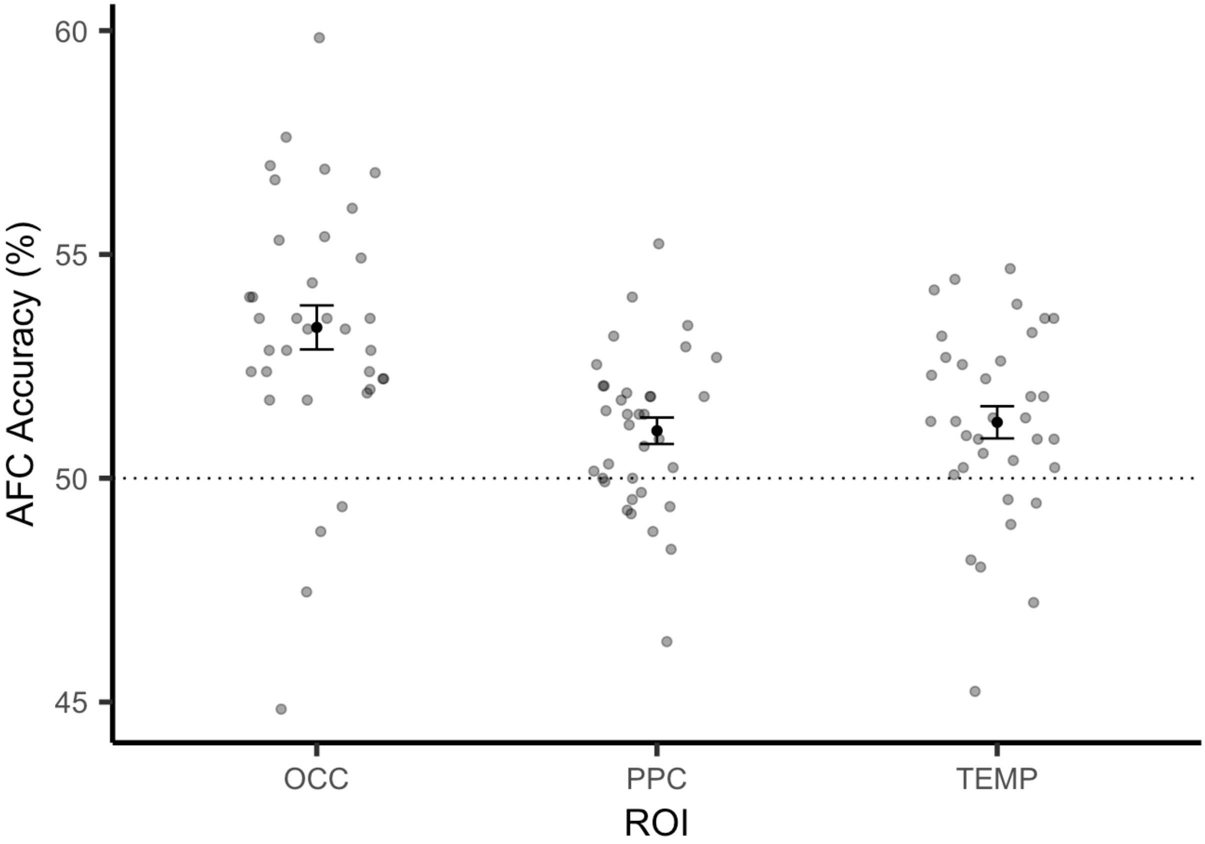

Figure 4.5. Alternative forced choice (AFC) accuracy for AAM components, modeled separately for occipital (OCC), posterior parietal (PPC), and temporal (TEMP) cortical ROIs. All three ROIs reconstructed face images at above chance levels, with OCC reconstructions performing significantly better than PPC and TEMP. Error bars represent SEM

The PPC region used included several different ROIs, we next sought to explore whether individual regions had predictive power (Fig. 4.6). We performed the same analysis, but focused on angular gyrus (ANG), intraparietal sulcus (IPS), supramarginal gyrus (SMG), and superior parietal cortex (SPC) separately (all four combined was our PPC ROI). Average performance in all four ROIs was similar to the PPC overall, indicating that there was little or no model gain from including all of these ROIs together. However, there was additional variance in model performance across participants (PPC overall: *SD* = 1.75; ANG: *SD* = 2.70; IPS: *SD* = 2.33; SMG: *SD* = 2.97; SPC: *SD* = 3.47). Thus when examined individually, only SPC performed significantly above chance (*t*(34) = 2.37, *p* = 0.02, *d* = 0.4, *M* = 51.4%). All other regions failed to reach significance at the 0.05 threshold (ANG: *t*(34) = 0.95, *p* = 0.35, *d* = 0.16, *M* = 50.4%; IPS: *t*(34) = 1.94, *p* = 0.06, *d* = 0.33, *M* = 50.8%; SMG: *t*(34) = 1.94, *p* = 0.06, *d* = 0.33, *M* = 51.0%). An overall ANOVA, however, indicated no significant difference between the 4 ROIs (*F*(3, 102) = 0.71, *p* = 0.55, η_G^2 = 0.01).

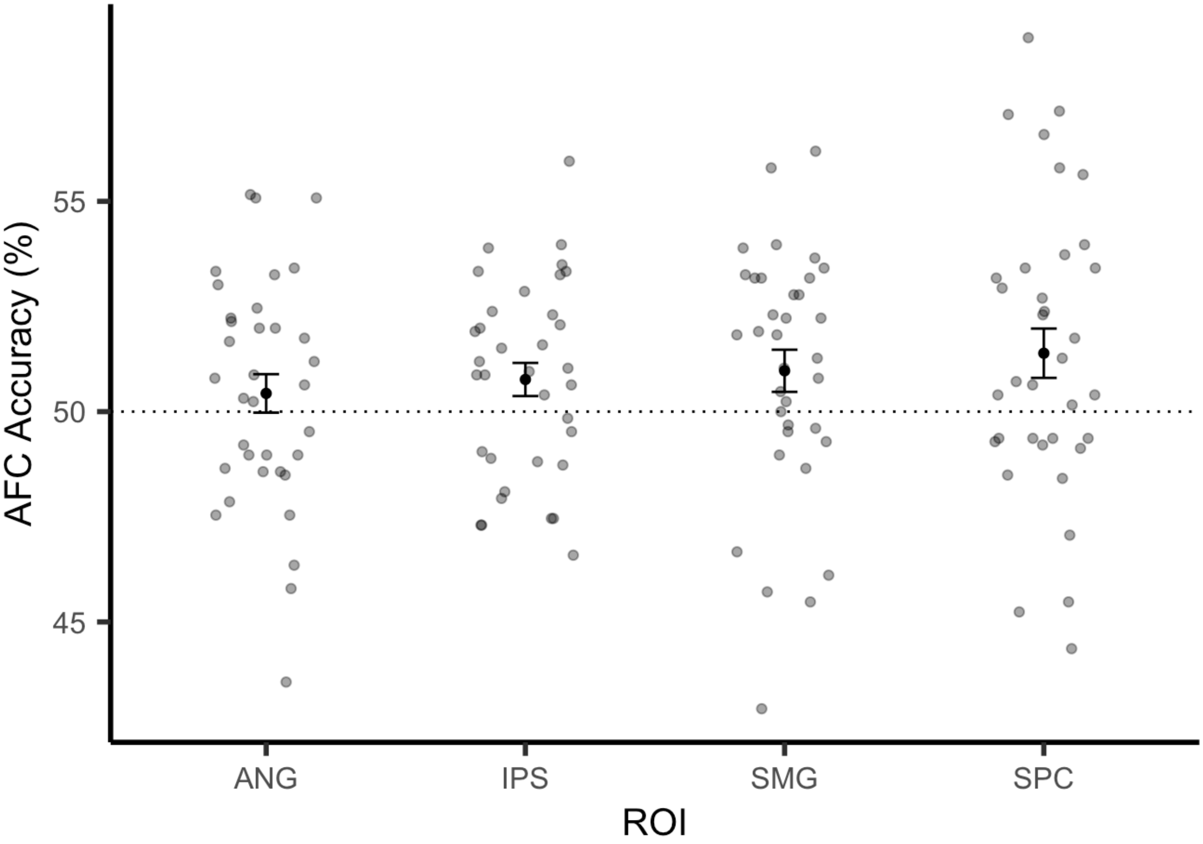

Figure 4.6. Alternative forced choice (AFC) accuracy for AAM components, modeled separately for ROIs within the overall PPC region: angular gyrus (ANG), intraparietal sulcus (IPS), supramarginal gyrus (SMG), and superior parietal cortex (SPC). Only predictions based on SPC activity performed significantly above chance. Error bars represent SEM

The TEMP region used included several different ROIs, we next sought to explore whether individual regions had predictive power (Fig. 4.7). We performed the same analysis, but focused on inferior temporal (Inf), superior temporal (Sup), middle temporal (Mid), temporal pole (Pole), and transverse temporal (Trans). We also included the fusiform gyrus (FUS), which was not included in the overall temporal ROI. Average performance was significantly above chance for Inf (t(34) = 3.35, p = 0.002, d = 0.57, M = 51.6%) and Sup (t(34) = 3.24, p = 0.003, d = 0.55, M = 51.3%). Mean performance was similar in FUS and Mid, but failed to reach significance (FUS: t(34) = 1.97, p = 0.056, d = 0.33, M = 51.0%; Mid: t(34) = 1.39, p = 0.17, d = 0.24, M = 50.8%). The mean performance was at chance levels for Pole and Trans (Pole: t(34) = 0.30, p = 0.77, d = 0.05, M = 50.1%; Trans: t(34) = 0.72, p = 0.47, d = 0.12, M = 50.2%). An overall ANOVA, however, indicated no significant difference between the 6 ROIs (F(5, 170) = 2.14, p = 0.063, η_G^2 = 0.04).

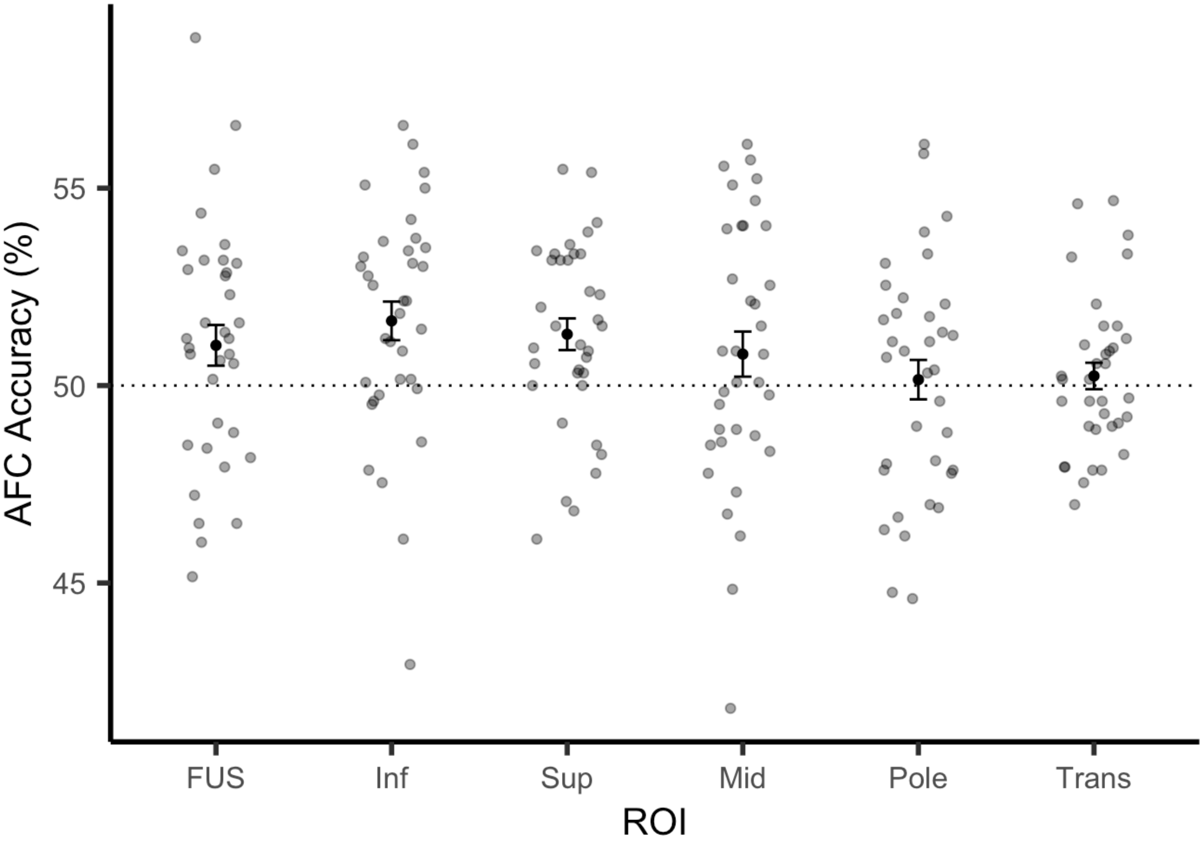

Figure 4.7. Alternative forced choice (AFC) accuracy for AAM components, modeled separately for ROIs within the broader temporal ROI: inferior temporal (Inf), superior temporal (Sup), middle temporal (Mid), temporal pole (Pole), and transverse temporal (Trans). We also included the fusiform gyrus (FUS), which was not included in the overall temporal ROI. Average performance was significantly above chance for Inf and Sup. Error bars represent SEM

Reconstruction performance compared to eigenface model

There was no evidence that 50 AAM components were more accurately reconstructed than the 300 eigenface components (Fig. 4.8; see Fig. 4.4 for a visualization). A model (AAM, eigenface) x ROI (OCC, PPC, TEMP) repeated measures ANOVA found no significant difference in model performance (*F*(1, 34) = 0.49, *p* = 0.49, η_G^2 = 0.001) and no interaction with ROI (*F*(2, 68) = 0.94, *p* = 0.39, η_G^2 = 0.001).

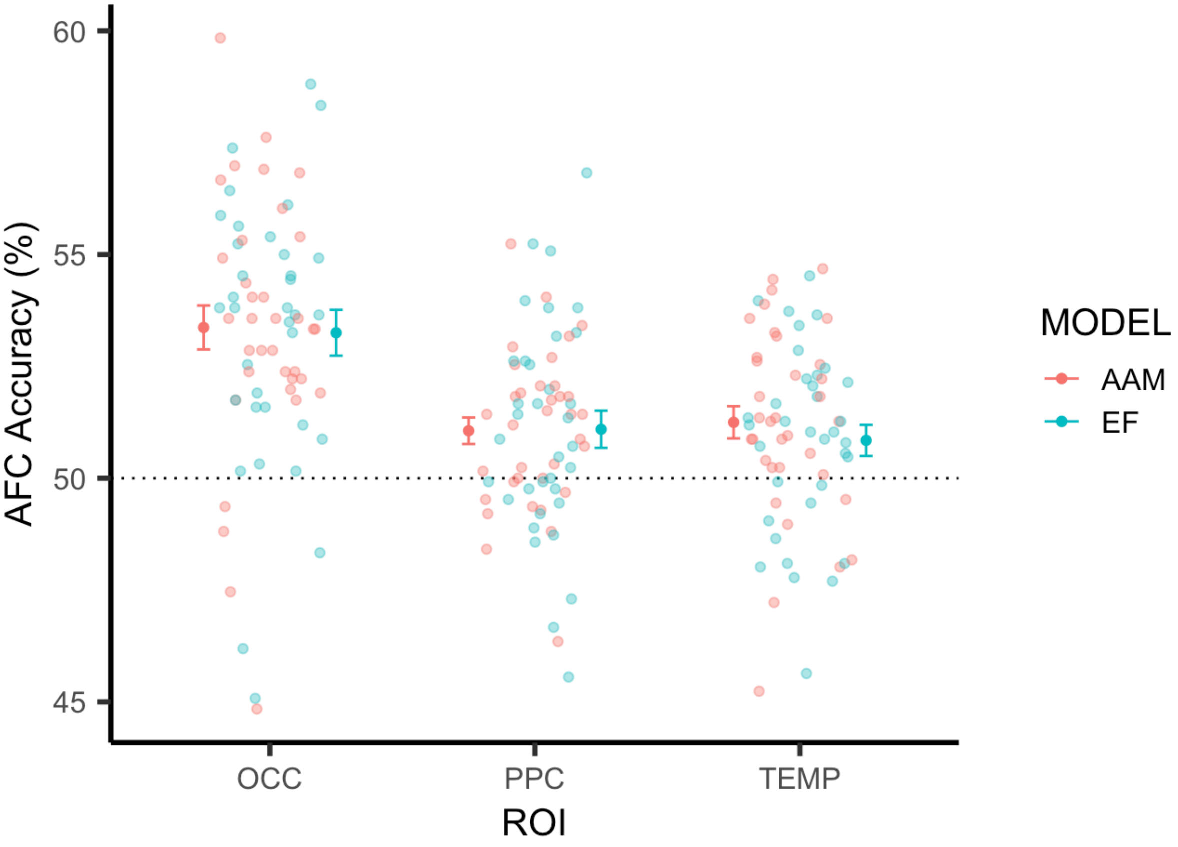

Figure 4.8. AFC accuracy for AAM compared to eigenface components. We found no difference in AFC accuracy comparing 50 AAM components (orange) to 300 eigenface (blue; EF) in occipital (OCC), posterior parietal (PPC), and temporal (TEMP) cortical ROIs. Error bars represent SEM

We followed this up with the same approach, but with our four individual PPC ROIS. We again found no significant difference in model performance (*F*(1, 34) = 0.64, *p* = 0.43, η_G^2 = 0.001) and no interaction (*F*(3, 102) = 0.48, *p* = 0.70, η_G^2 = 0.002). We repeated this approach with the six TEMP ROIs and again found no significant difference in model performance (*F*(1, 34) = 0.09, *p* = 0.77, η_G^2 < 0.001) and no interaction (*F*(5, 170) = 1.13, *p* = 0.34, η_G^2 = 0.004).

Although there was no overall difference found, this may have been an unfair test due to differences in the number of components included. Therefore, we repeated our model (AAM, eigenface) x ROI (OCC, PPC, TEMP) repeated measures ANOVA, but with only the top 50 eigenface components. We found no significant difference in model performance (*F*(1, 34) = 0.33, *p* = 0.57, η_G^2 < 0.001) and no interaction (*F*(2, 68) = 1.27, *p* = 0.29, η_G^2 = 0.002). We further examined whether there were any accuracy differences as different numbers of components were included in the analysis (Fig. 4.9). For the purposes of statistical analysis, we looked at 2, 4, 6, 8, and 10 components included. We proceeded in steps of 2 so that both 1 additional shape and 1 additional appearance component could be added at each step (ordered in terms of variance added). We limited ourselves to the first few components in order to avoid overfitting the model. We ran a series of model (AAM, eigenface) x component number (2, 4 ,6, 8 ,10) repeated measures ANOVAs for each ROI (OCC, PPC, TEMP). In OCC, we found a significant main effect of component number, indicating an improvement in performance with more components included (*F*(4, 136) = 23.20, *p* < 0.001, η_G^2 = 0.03). Consistent with our previous analyses, there was no significant effect of model (*F*(1, 34) = 1.39, *p* = 0.25, η_G^2 = 0.004). There was however, a significant interaction between component number and model (*F*(4, 136) = 2.54, *p* = 0.04, η_G^2 = 0.003). This seems to indicate some differences in the rate of AFC improvement for each model as more components are added, but these differences even out over time. We found no significant effects for PPC or TEMP.

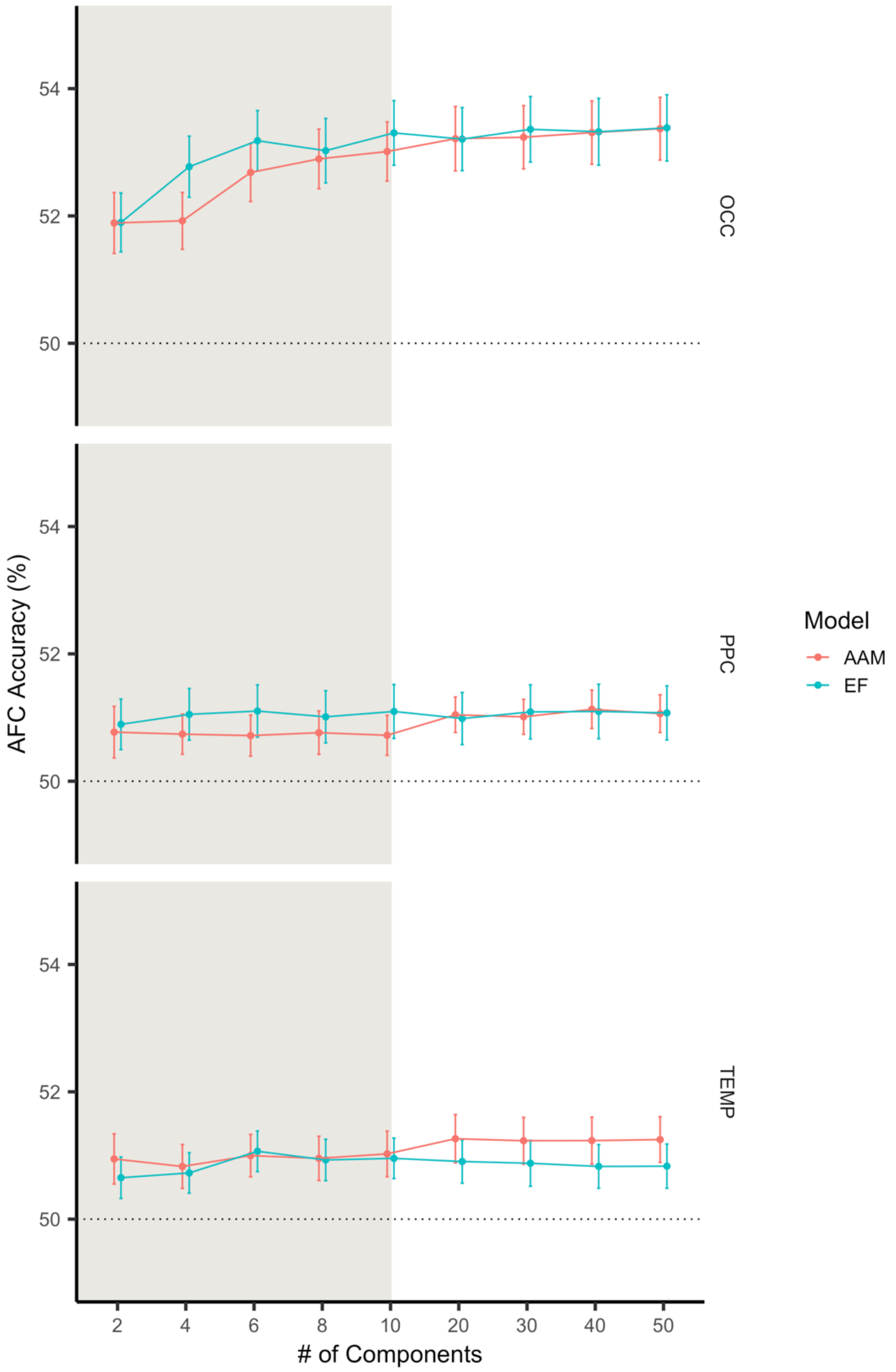

Figure 4.9. AFC accuracy by the number of components included for AAM (orange) and EF (blue) components. In OCC (top), performance improved at differential rates over the first 10 components, but plateaued at the same level for both. In PPC (middle) and TEMP (bottom), both plateaued immediately for both model types. The components included are ordered in terms of explained visual variance for each model type. Only even numbers were included, so that for the AAM components, each step included the next top shape and the next top appearance component. Shaded region indicates zoomed in region of graph that shows transitions of 2 components, rather than 10 on the rest of the x-axis. Error bars represent SEM

Reconstruction performance for appearance vs shape components

One of the advantages of using AAM, is that the components can be divided into two broad categories. This allows us to look at differences between regions in terms their ability to reconstruct different types of components (Fig. 4.10). Model performance was assessed in the same way here, but with the components divided into 25 shape and 25 appearance components. A model (shape, appearance) x ROI (OCC, PPC, TEMP) repeated measures ANOVA found no significant difference in model performance (*F*(1, 34) = 3.12, *p* = 0.09, η_G^2 = 0.02). However, there was a significant interaction (*F*(2, 68) = 10.88, $p < 0.01$, η_G^2 = 0.03). Follow up t-tests found no difference comparing appearance and shape for OCC (*t*(34) = 1.06, *p* = 0.30, *d* = 0.17). In TEMP, however, the model performed significantly better at predicting shape than appearance components (*t*(34) = 3.76, $p < 0.001$, *d* = 0.80). PPC trended in that same direction, but did not quite reach significance (*t*(34) = 1.96, *p* = 0.058, *d* = 0.45). Follow-up one sample t-tests found that AFC performance for shape was significantly above chance for PPC (*t*(34) = 4.42, $p < 0.001$, *d* = 0.75) and TEMP (*t*(34) = 5.82, $p < 0.001$, *d* = 0.98). However, both failed to reach significance for appearance (PPC: *t*(34) = 1.22, *p* = 0.23, *d* = 0.21; TEMP: *t*(34) = 1.30, *p* = 0.20, *d* = 0.22).

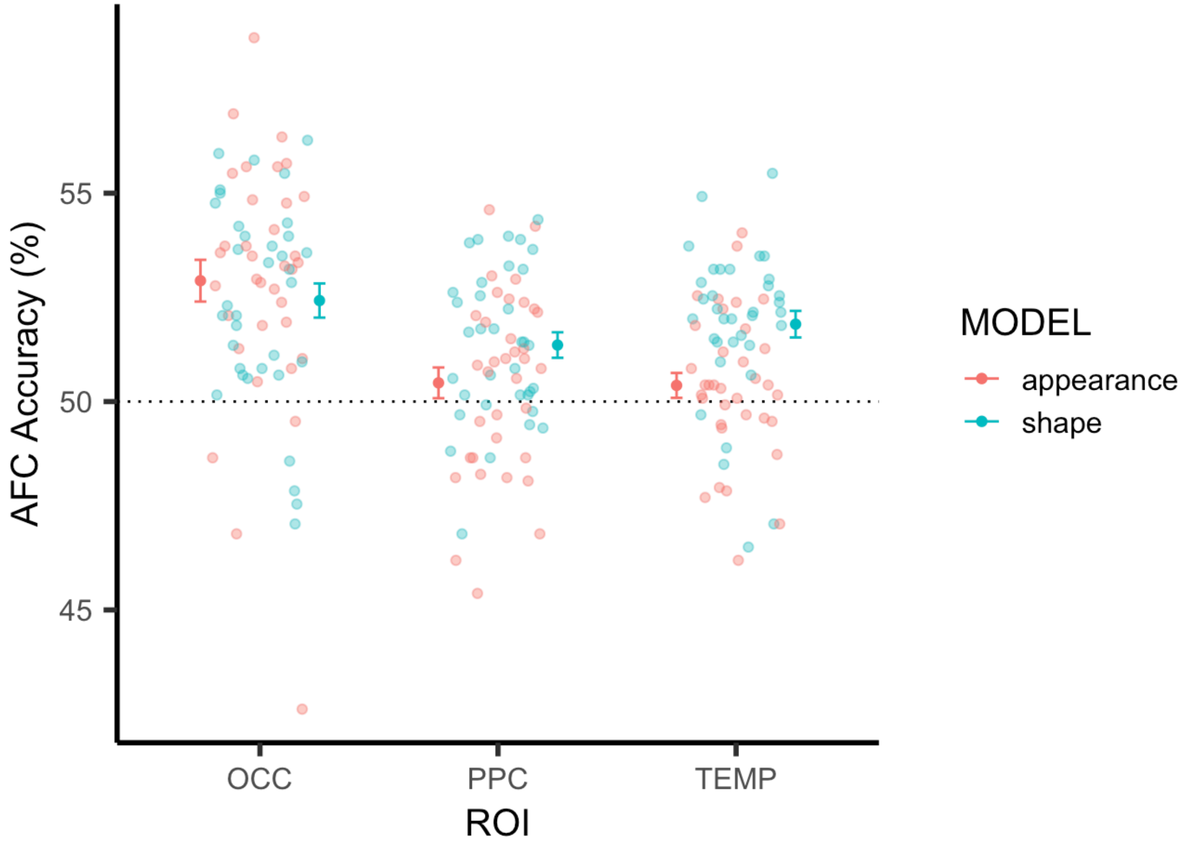

Figure 4.10. AFC accuracy for shape compared to appearance components. AFC accuracy was significantly higher for the 25 shape (blue) than the 25 appearance (orange) AAM components for TEMP. There was a trend in the same direction for PPC, but no evidence for a difference in OCC. Error bars represent SEM

We found a trend towards an overall advantage in PPC for shape compared to appearance components. We followed-up by looking at ROIs within PPC (Fig. 4.11). A model (appearance, shape) x ROI (ANG, IPS, SMG, SPC) repeated measures ANOVA found a significant difference in model performance, with shape performing significantly better than appearance across these ROIs (*F*(1, 34) = 7.83, *p* = 0.01, η_G^2 = 0.03) and no significant interaction (*F*(3, 102) = 2.14, *p* = 0.10, η_G^2 = 0.02). Although there was no significant interaction, follow-up t-tests revealed significantly better performance for shape components in ANG (*t*(34) = 2.85, *p* = 0.01, *d* = 0.63) and SMG (*t*(34) = 2.51, *p* = 0.02, *d* =-0.58), but not for IPS (*t*(34) = 0.76, *p* = 0.45, *d* = 0.20) or SPC (*t*(34) = 0.26, *p* = 0.79, *d* = 0.05. In fact, AFC performance for shape was significantly above chance for ANG (*t*(34) = 2.67, *p* = 0.01, *d* = 0.29) and SMG (*t*(34) = 3.26, *p* = 0.003, *d* = 0.55), but both failed to reach significance for appearance (PPC: *t*(34) = 0.92, *p* = 0.36, *d* = 0.16; SMG: *t*(34) = 0.06, *p* = 0.95, *d* = 0.01).

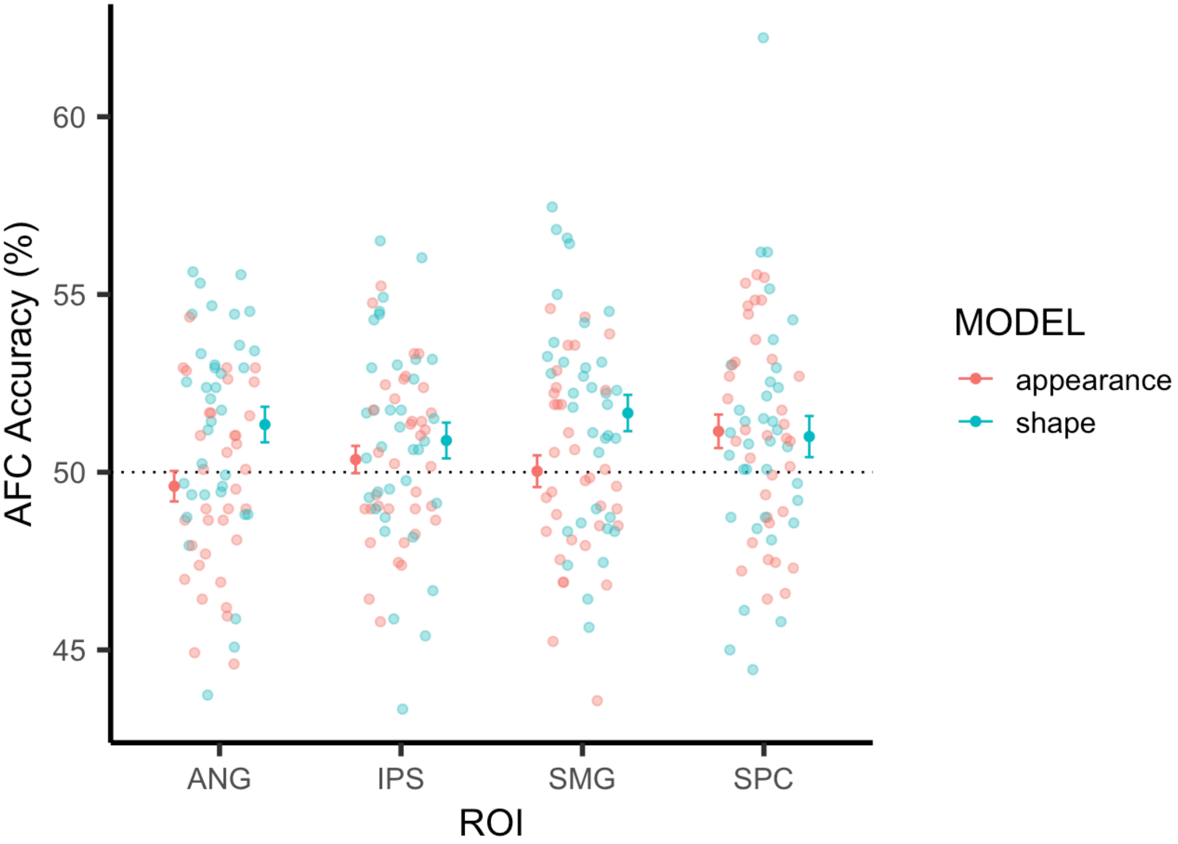

Figure 4.11. AFC accuracy for shape compared to appearance components within PPC ROIs. AFC accuracy was significantly higher for the 25 shape (blue) than the 25 appearance (orange) AAM components for two PPC ROIs: ANG and SMG, but no different for IPS and SPC. Error bars represent SEM

We found an overall advantage in temporal cortex for shape compared to appearance components. We followed-up by considering TEMP subregions (Fig. 4.12). A model (appearance, shape) x ROI repeated measures ANOVA found a significant difference in model performance, with again, shape performing significantly better than appearance across these ROIs (*F*(1, 34) = 7.75, *p* = 0.01, η_G^2 = 0.03) and no significant interaction (*F*(5, 170) = 1.30, *p* = 0.27, η_G^2 = 0.01) or effect of ROI (*F*(5, 170) = 1.89, *p* = 0.1, η_G^2 = 0.02). Although there was no significant interaction, follow-up t-tests revealed significantly better performance for shape components in Sup ($t(34) = 3.73$, $p < 0.001$, $d = 0.72$) and Mid ($t(34) = 2.88$, $p = 0.01$, $d = 0.60$), but not for the other included regions. In fact, performance for shape was significantly above chance for Sup ($t(34) = 5.36$, $p < 0.001$, $d = 0.91$) and Mid ($t(34) = 3.61$, $p < 0.001$, $d = 0.61$), but both failed to reach significance for appearance (Sup: $t(34) = 0.92$, $p = 0.36$, $d = 0.16$; Mid: $t(34) = 0.07$, $p = 0.94$, $d = 0.01$).

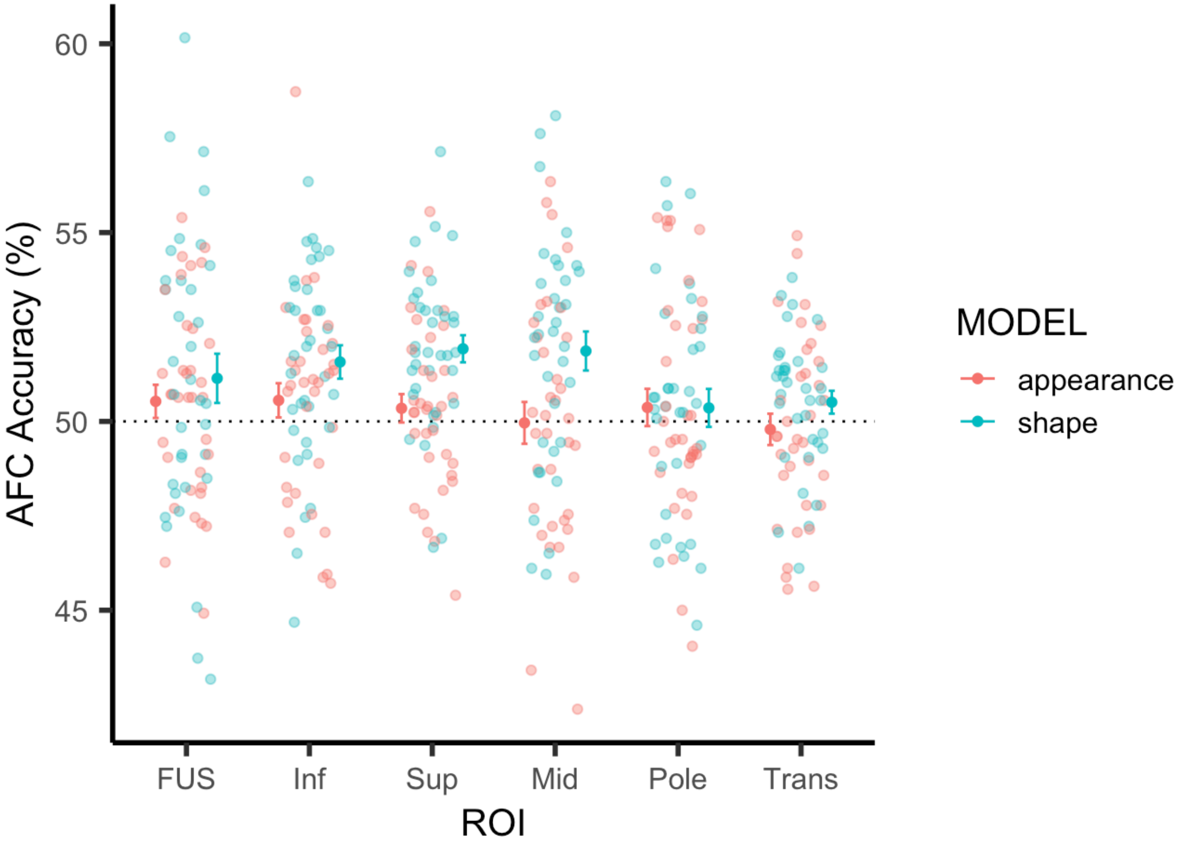

Figure 4.12. AFC accuracy for shape compared to appearance components within temporal ROIs. AFC accuracy was significantly higher for the 25 shape (blue) than the 25 appearance (orange) AAM components for two TEMP ROIs: Sup and Mid. There were no other significant differences in the included ROIs. Error bars represent SEM

Reconstruction performance for individual AAM components

Although accuracy for shape and appearance components provides an overall summary of what is best reconstructed from fMRI activity, the limited number of components (50) gives the additional opportunity to examine reconstruction performance for individual components. For each participant, we calculated the correlation between the predicted and true score across all 36 test images (Fig. 4.13). For the purpose of statistical testing, we converted all correlations into Fisher's Z and ran a series of one sample t-tests compared to a correlation of 0. Performance in OCC was significantly better than chance for 10 out of 25 shape components and 12 out of 25 appearance components (22/50 total). The pattern of significance was biased towards early components, but not exclusively so. For PPC, 5 out of 25 shape components were significantly predicted and 4 out of 25 appearance components. Interestingly, although the significant appearance components were biased towards early ones (3, 4, 5, 12), there did not appear to be the same pattern in shape components (1, 10, 17, 18, 20). For TEMP, 6 out of 25 shape components were significantly predicted and 3 out of 25 appearance components. Interestingly, there was a high degree of correspondence between PPC and TEMP, with all 5 significant shape components for PPC included in the 6 total for TEMP, and all 3 significant appearance components for TEMP included in the 4 significant for PPC. Although many more components were significant for OCC, the pattern aligned with TEMP and PPC, with only appearance component 12 being significantly predicted by only PPC. There was a total of 3 appearance components (A3, A4, A5) and 5 shape components (S1, S10, S17, S18, S20) significant for all three ROIs. Of note, although the most consistently predicted components were high in terms of appearance variance explained, that same pattern does not appear to hold for shape, where many later components were the best predicted across these ROIs (see Fig. 4.14 for a visualization of these components).

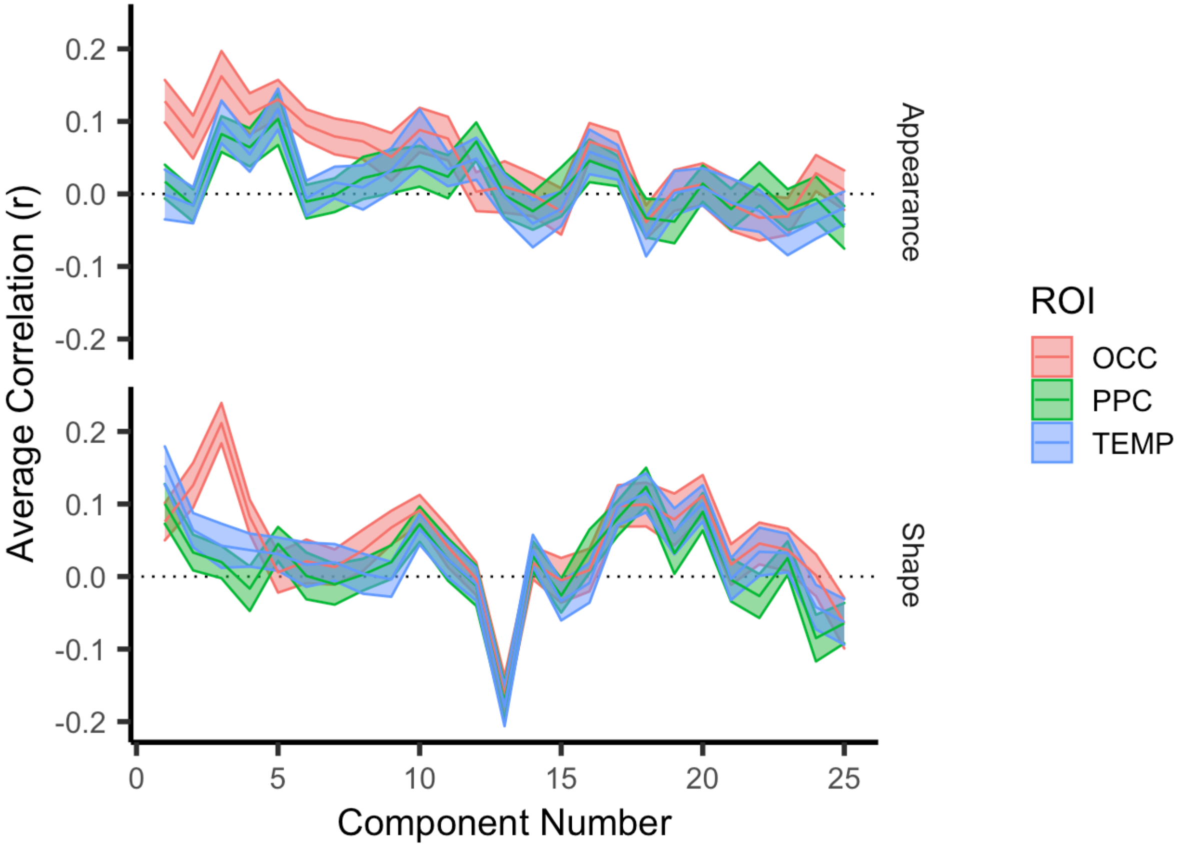

Figure 4.13. Average correlation between predicted and true score on individual AAM components, sorted by 25 appearance (top) and 25 shape (bottom) components, and by OCC (red), PPC (green), and TEMP (blue). The components are ordered in terms of visual variance explained. Shaded areas represent SEM

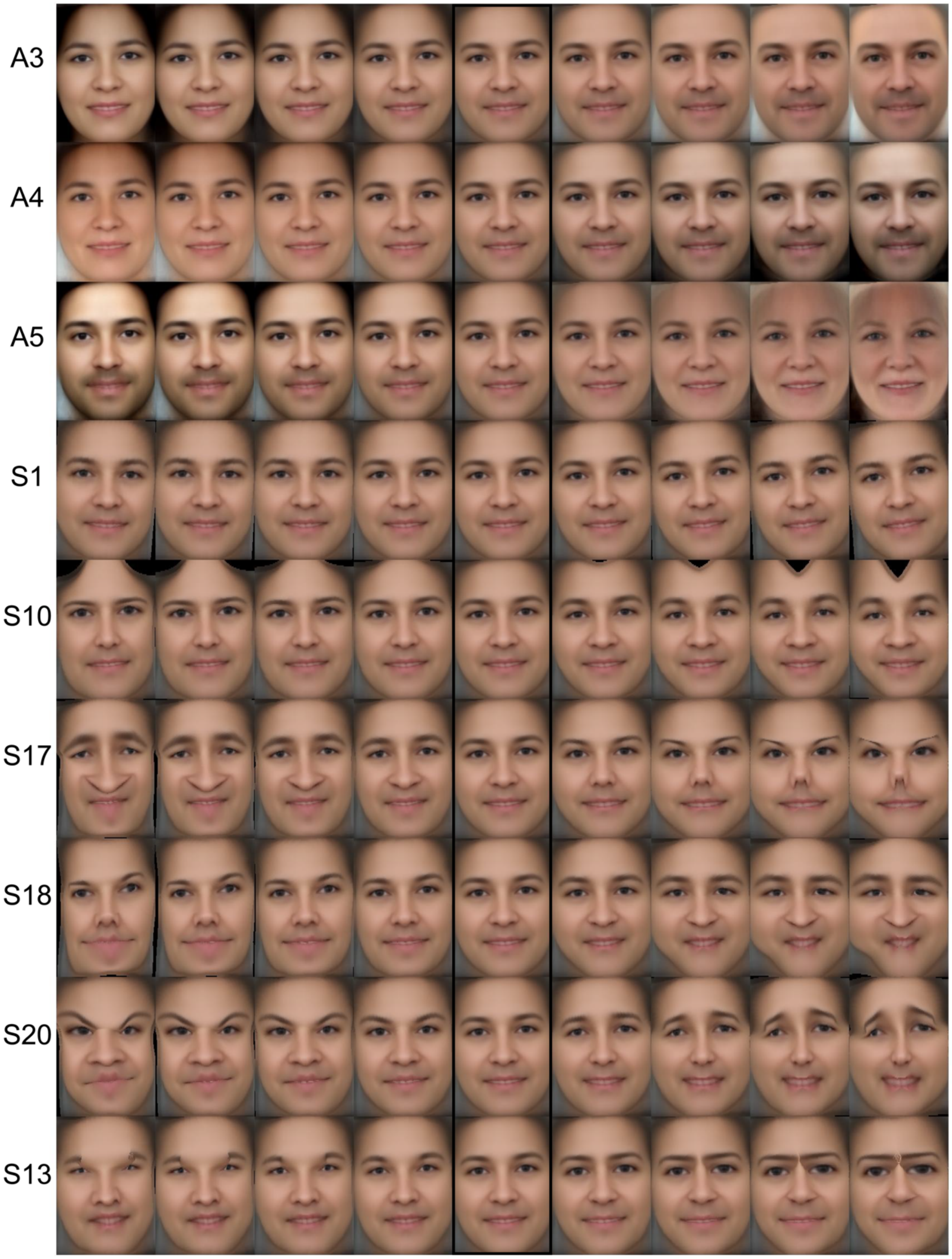

Figure 4.14. Best (and worst) predicted AAM components. The mean face (center) is depicted, shifted uniform amounts for each component (rows). The selected components were significantly predicted by all 3 ROIs (OCC, PPC, TEMP). There were 3 appearance components (A3, A4, A5) that met this criteria and 5 shape components (S1, S10, S17, S18, S20). One additional shape component (S13) is also pictured as the predictions were significantly negatively correlated with the true values.

Interestingly, there were a few components with a significantly negative correlation (7 total: 1 OCC shape, 3 PPC shape, 1 TEMP shape, 2 TEMP appearance; see Fig. 4.14, S4.6). Of particular note, shape component 13 had a negative correlation for all 3 ROIS. A subsequent investigation of this unexpected effect, found that it was driven by a small number of stimuli (5) which were outliers in terms of model predictions on this component, all in the negative direction. When those stimuli are removed from the analysis, predictions for this component are no longer negatively correlated with the true scores (OCC: *M* = 0.05 ± 0.18; PPC: *M* = 0.05 ± 0.17; TEMP: *M* = 0.07 ± 0.17).

The effect of repetition on reconstruction performance

For our initial set of analyses, we assumed that averaging across repetitions of test faces would yield the strongest reconstruction performance. However, we proceeded to examine the test trials separately, in order to both test that assumption and to examine whether there were any meaningful differences between the first and second time a stimulus is seen (Fig. 4.15). For ease of interpretation, for this analysis, we focused on participants (N=23) who only saw the test items twice (see Methods). A repetition (1, 2) x ROI (OCC, PPC, TEMP) repeated measures ANOVA found no significant difference in repetition (*F*(1, 22) = 3.78, *p* = 0.06, η_G^2 = 0.04) and no interaction (*F*(2, 44) = 2.45, *p* = 0.10, η_G^2 = 0.01). However, follow up t-tests revealed a significant drop in performance between repetitions 1 and 2 in OCC (*t*(22) = 3.12, *p* = 0.005, *d* = 0.65), but not in PPC (*t*(22) = 1.33, *p* = 0.20, *d* = 0.39) or TEMP (*t*(22) = 0.61, *p* = 0.55, *d* = 0.18).

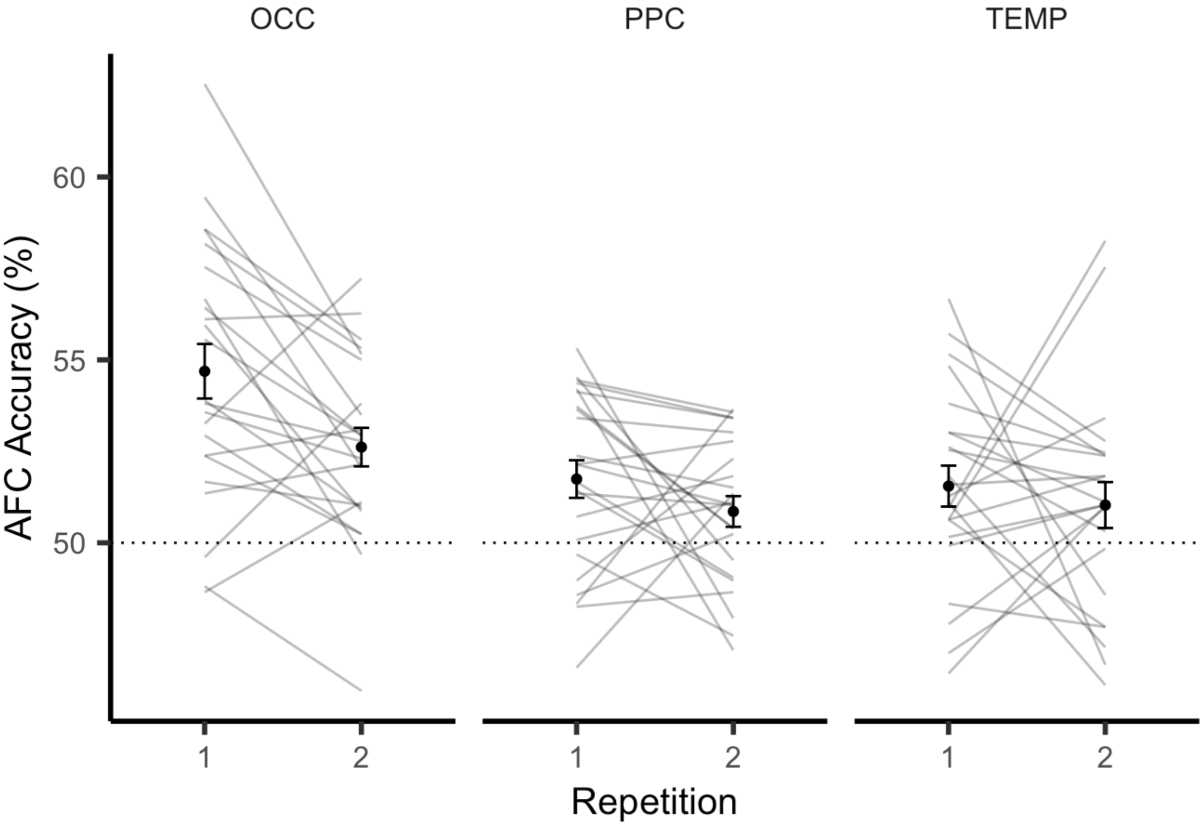

Figure 4.15. AFC accuracy by repetition number. There was a significant drop in AFC accuracy for predictions made by OCC on the 2nd test trial compared to the 1st test trial of each test image. There was no effect of repetition in PPC or TEMP. For ease of interpretation, the current plot and results focus on participants (N=23) who saw all test items only twice. Error bars represent SEM

As a follow-up, separating the predictions allowed for looking at whether there was any gain in performance on correct compared to incorrect trials. Participants were consistently accurate at the task, so there were too few incorrect trials to examine separately. Instead, we reran the same analyses on correct trials only and found the same pattern of statistical results (see Fig. S4.1).

We were further interested in whether there would be any pattern within PPC (Fig. 4.16). However, a repetition (1, 2) x ROI repeated measures ANOVA found no significant difference in repetition (*F*(1, 22) = 1.54, *p* = 0.23, η_G^2 = 0.01) and no interaction (*F*(3, 66) = 1.37, *p* = 0.26, η_G^2 = 0.02). Follow up t-tests revealed no significant differences in any ROI between repetition 1 and 2. Follow-up analyses looking at correct trials only confirmed this same pattern of results (see Fig. S4.2).

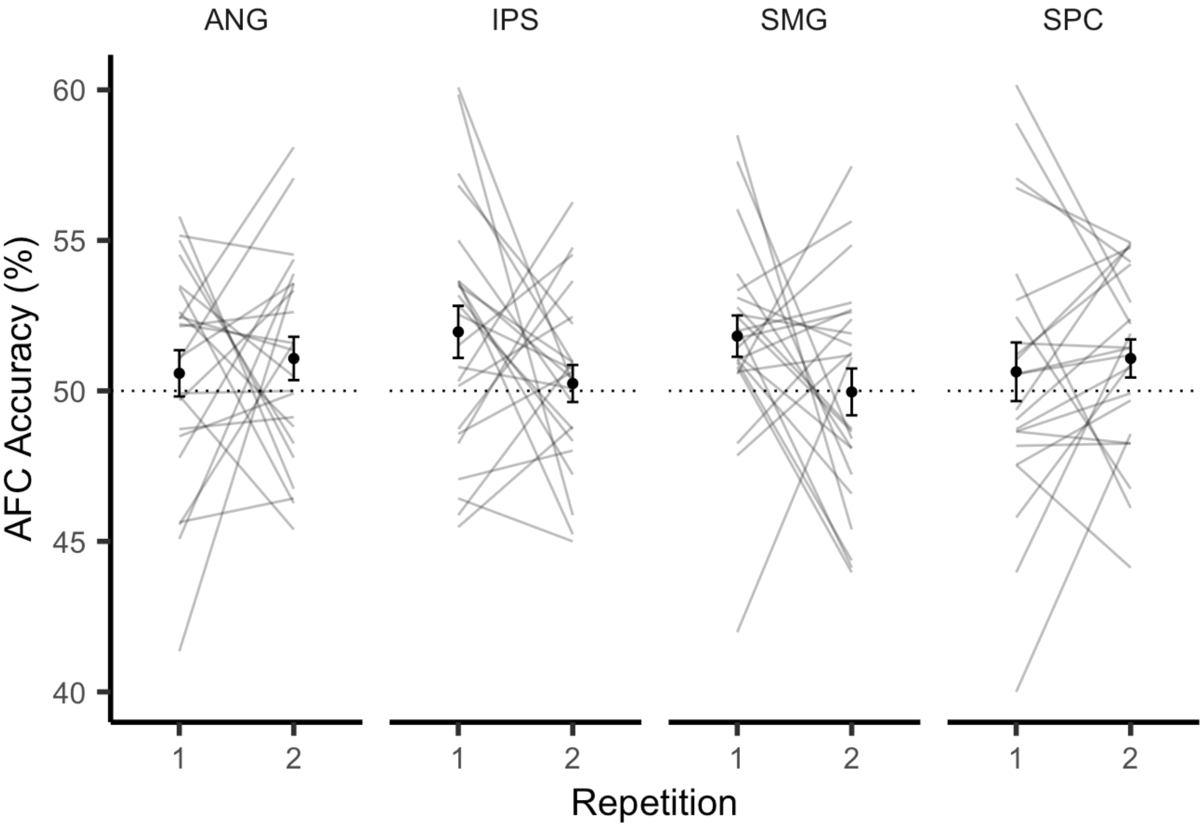

Figure 4.16. AFC accuracy by repetition number for PPC ROIs. There was no significant effect of repetition on AFC accuracy for any individual PPC ROI (ANG, IPS, SMG, SPC). For ease of interpretation, the current plot and results focus on participants (N=23) who saw all test items only twice. Error bars represent SEM

We were also interested in whether there would be any pattern within TEMP (Fig. 4.17). However, a repetition (1, 2) x ROI repeated measures ANOVA found no significant difference in repetition ($F(1, 22) = 2.77$, $p = 0.11$, $\eta_G^2 = 0.03$) and no interaction ($F(5, 110) = 0.67$, $p = 0.65$, $\eta_G^2 = 0.01$). Follow up t-tests revealed no significant differences in any ROI between repetition 1 and 2. Follow-up analyses looking at correct trials only broadly were in line with this pattern of results. (see Fig. S4.3). In this case, however, FUS which had trended towards significance

(t(22) = 2.05, p = 0.053, d = 0.59), did significantly differ, with performance falling on the repeated trial (t(22) = 2.52, p = 0.020, d = 0.59).

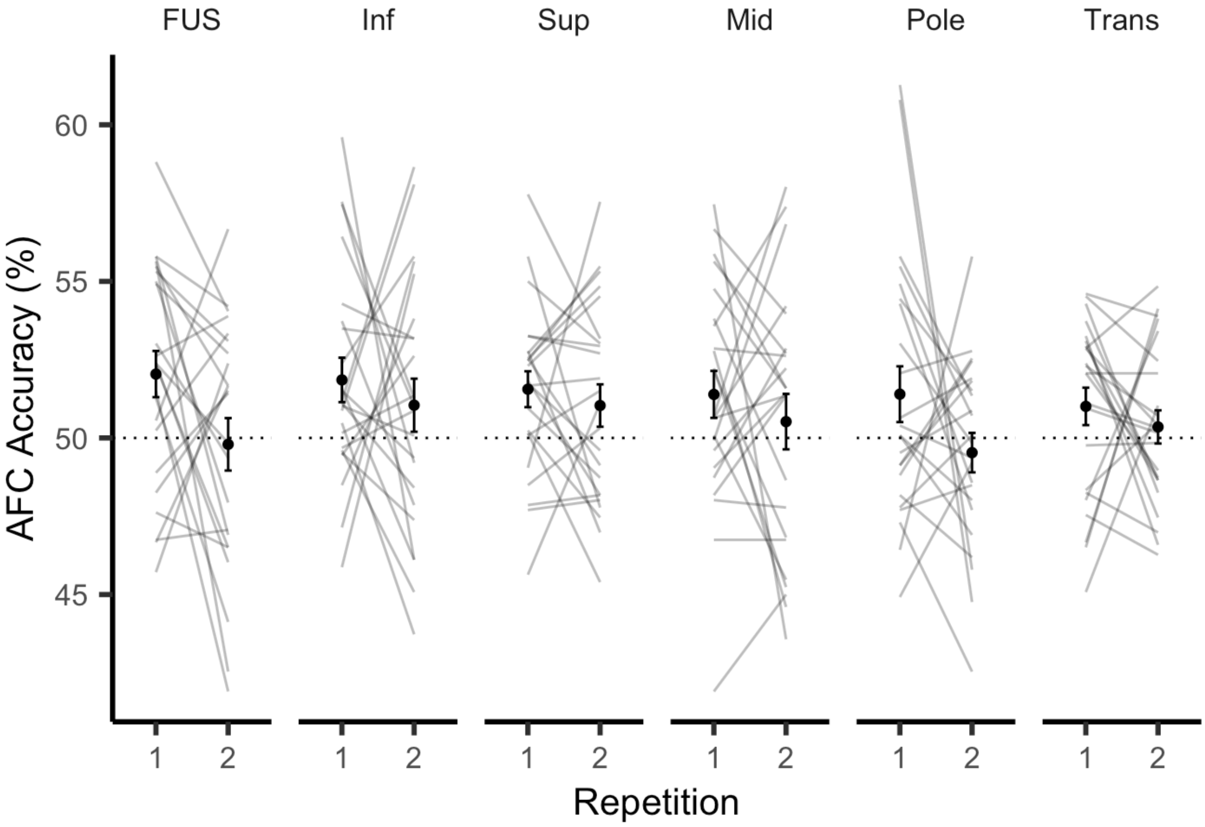

Figure 4.17. AFC accuracy by repetition number for temporal ROIs. There was no significant effect of repetition on AFC accuracy for any individual TEMP ROI (FUS, Inf, Sup, Mid, Pole, Trans). For ease of interpretation, the current plot and results focus on participants (N=23) who saw all test items only twice. Error bars represent SEM

Predicting subjective ratings

In order to see how well perceptually-important information can be predicted, we used the same modelling procedure, but replaced the AAM components with five subjectively-rated dimensions (affect, gender, trustworthiness, dominance, and attractiveness). We used the same procedure as with individual AAM components to calculate the relationship between predicted and true subjective rating values (Fig. 4.18). We found that model performance was significantly above chance for trustworthiness in all three ROIs (OCC: t(34) = 5.47, p <

0.001, *d* = 0.93, *M* = 0.13; PPC: *t*(34) = 4.15, *p* < 0.001, *d* = 0.7, *M* = 0.10; TEMP: *t*(34) = 4.55, *p* < 0.001, *d* = 0.77, *M* = 0.11). Performance was above chance for affect (*t*(34) = 7.04, *p* < 0.001, *d* = 1.19, *M* = 0.19) and gender (*t*(34) = 3.60, *p* = 0.001, *d* = 0.61, *M* = 0.12) only for OCC. Dominance and attractiveness predictions were not significantly correlated with the true values in any ROI.

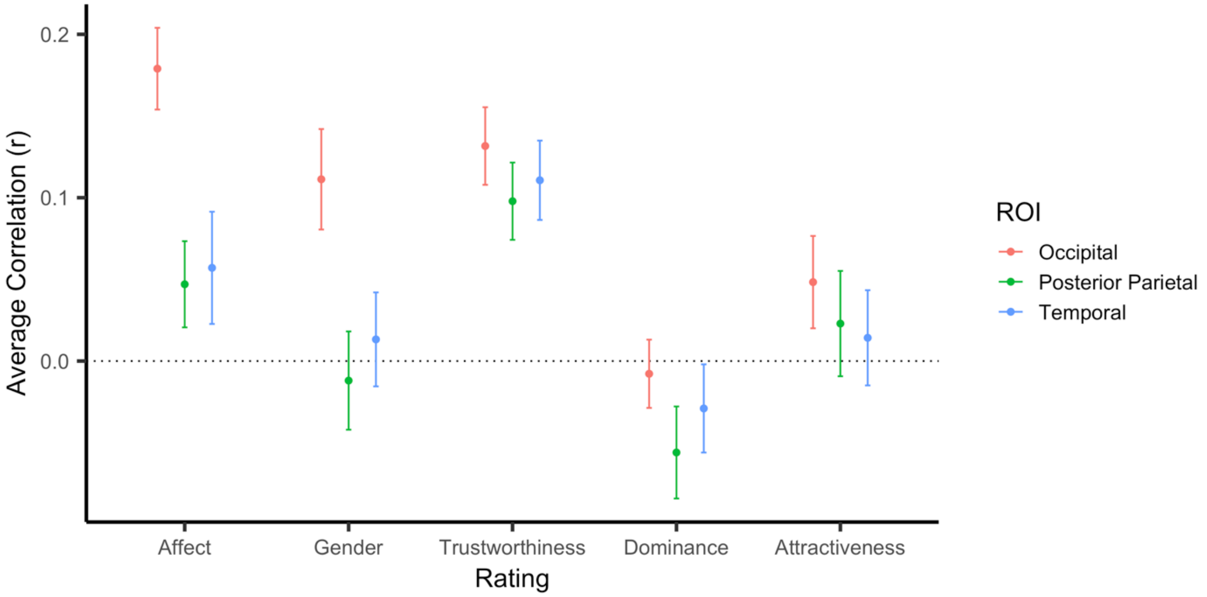

Figure 4.18. Average correlation (r) between predicted and averaged subjective ratings (affect, gender, trustworthiness, dominance, and attractiveness) for OCC (red), PPC (green), and TEMP (blue). For plotting purposes, we used r, but for statistical testing, we converted the correlations to Fisher's z. Model performance was significantly above chance for trustworthiness in all three ROIs, and above chance for gender and affect only in OCC. Error bars represent SEM

Discussion

The current study evaluated the ability to reconstruct face images based on distributed patterns of fMRI activity evoked during the perception of the face stimuli. We focused on a data-driven approach to parameterizing face images, the active appearance model (AAM). We then used ridge regression to predict the top 50 AAM components based on fMRI activity within different ROIs. We evaluated the performance of these models with alternate force choice (AFC) accuracy in relation to the true component values. This analysis established the model's ability to reconstruct these face components at above chance levels across several different cortical ROIs. Further, we established differences in the ability to predict certain types of components within particular ROIs.

Differences in content representation

One advantage of the AAM is that it allows for the comparison of two broad classes of features (see Chapter 2). The shape components are derived first and come from the top 25 principal components of the locations of several face landmarks (e.g. landmarks along the chin, around the eyes). These components represent a mix of holistic (e.g. facial expression), configural (e.g. relative position of eyes), and local information (e.g. nose or mouth shape). They also capture less perceptually-important information, such as head-tilt. The appearance components are the top 25 principal components once all shape variance is removed. These components represent all information related to color and texture, including perceptually-important (e.g. facial hair) and unimportant (e.g. lighting) information. Broadly, the shape components may tend to represent more high-level information (e.g. affect), whereas appearance components capture more low-level visual information (e.g. skin tone).

We found no difference in the ability to reconstruct either component type in occipital cortex. Further investigation could look at regions within our broad occipital ROI to identify whether there were any areas that performed better at predicting one component type. In contrast to occipital, we found that the temporal cortex better reconstructed shape than appearance information. This is unsurprising, given that it is later in the visual processing stream and tends to represent higher-level information (Cichy et al., 2014; Martin et al., 2018). When we looked at regions within posterior parietal and temporal cortex, we found that when there was a difference in shape and appearance, that shape was better predicted. In particular, within PPC, we found this pattern in ANG and SMG, and within temporal, we found this pattern in Sup and Mid. The results in ANG are of particular interest here because of its role in representing the content of memory (Kuhl & Chun, 2014). This region also previously performed best at reconstructing face images from working memory (Lee & Kuhl, 2016). The current analysis suggests this was likely driven by higher-level visual features.

We further examined how well the model performed at predicting individual components. We found that the components that are best predicted by fMRI patterns do not necessarily correspond to visual variance explained. This was especially true for the shape components, where the components most consistently predicted across the three ROIs tended towards lower visual importance (components 10, 17, 18, 20). For the appearance components, there does appear to be a relationship between better predictions and higher visually important components, however it is not a completely linear pattern, with some later components (10, 11, 16 ,17) being significantly predicted by occipital cortex.

Examining model performance in terms of individual components helps to explain similarities and differences between the ROIs. Our analyses identified three appearance components that were consistently predicted by all three main ROIs. All three of those components appear to partially capture variance associated with gender, and to a lesser extent skin tone. They also appear to capture less perceptually-important information (e.g. lighting). Our analyses also identified five shape components that were consistently predicted by all three ROIs. Many of these components reflect complex features relevant to the specific identity of faces, such as nose shape and size. Many of these components also capture eyebrow position, which is likely related to higher-level information, such as affect. Outside of these components, there were several components that were predicted only by one or two of the ROIs. These components have a wide range of potential interpretations and span the range of components (see Fig. S4.4,5 for a visualization of the components).

As one way to provide additional insight to what the model may be decoding, we utilized the same modelling approach to decode perceptually-important dimensions across the same brain regions (affect, gender, dominance, trustworthiness, and attractiveness). Here we found that OCC significantly predicted affect, gender, and trustworthiness. PPC and TEMP both only significantly predicted trustworthiness. Given the importance of trustworthiness in immediate perceptual evaluations (Oosterhof & Todorov, 2008) and previous evidence of its representation in the brain (Cao et al., 2020), it is not surprising that it was one of the best predicted subjective dimensions. We, however, did expect to be able to successfully predict more of these dimensions. In particular, the failure of PPC and TEMP to predict gender is surprising given the discussion above about how some of the best predicted individual AAM components appear to load strongly onto gender.

Although we have evidence that multiple AAM components and subjective ratings can be decoded from neural activity, this analysis does not establish what is driving the success of the model. For example, certain neurons could be tuned to a particular AAM component, to another dimension related to a component (e.g. gender), to a face exemplar high or low on that component, or to one particular feature of a component (e.g. nose shape). Furthermore, regions that successfully predict the same component, may not be driven by the same underlying neural representation.

Effect of repetition

Although we expected that utilizing the average voxel activity pattern across two test trials to yield the best results, we were also interested in whether there was any difference in the ability to reconstruct images based on the first or second time a test image was presented. We found evidence for a drop in performance between repetitions 1 and 2 in occipital cortex and fusiform gyrus, though the latter was only significant when incorrect trials were removed. Although we were specifically interested in parietal regions that are involved in memory, we failed to find any regions that demonstrated significant evidence of “memory amplification”, where the second appearance of the item was better represented (Favila et al., 2018).

There are several potential explanations for why we may have seen the drop in performance for the repeated trial. One explanation is that the first time a face was seen demanded more attention, because participants were encoding the features of the faces. On the second appearance, however, participants may have gotten an immediate familiarity signal and responded with “old” and proceeded to lapse in attention for the remainder of the trial. One alternative account is that the model was trained on only the first appearance of face images, therefore any differences in repetition trials, whether due to an attention or memory effect, were not trained on. There are too few repetition trials to explore this possibility with the present data.

Comparison to eigenfaces

Previous efforts to reconstruct faces have successfully employed eigenfaces to parameterizing face images (Cowen et al., 2014; Lee & Kuhl, 2016). Here we used an improved parameterization technique that represents face images more realistically and efficiently, and has previously demonstrated greater reconstruction success (Chang & Tsao, 2017). Despite our expectations, we found no advantage to using the AAM compared to eigenfaces.

One potential explanation was differences in the number of components used (300 for eigenface, 50 for AAM). However, we found no advantage for 50 AAM components when compared to 50 eigenface components. Furthermore, although our choice of 50 AAM components was based on previous usages (Chang & Tsao, 2017) and not our own data, we found no evidence that a different number of components would have been better.

An important consideration is that although the AAM approach was no better than eigenface when measured with AFC accuracy, this approach to assessing performance is only expressed in relation to the predictability of the components themselves. That is, although both methods performed about as well at reconstructing within that space, the AAM renders more realistic reconstructions, which could potentially be judged as more similar to the true face based on subjective judgments. A behavioral study would need to asses that possibility.

Future directions

Although the present results establish the ability to reconstruct face images at above chance levels, the magnitude of the effect was modest compared to recent attempts that took a similar methodological approach (Cowen et al., 2014; Lee & Kuhl, 2016) or to other recent approaches with strong reconstruction performances (Dado et al., 2022; Güçlütürk et al., 2017; VanRullen & Reddy, 2019). Moving forward, this is a rich dataset that needs to be explored further in order to increase the power and effectiveness of this approach.

Our ultimate goal here, was not only to maximize face reconstruction accuracy, but to do so in a way that was practical to implement as part of a larger experiment where internal face representations could be a decodable dependent variable. For this reason, we focused on regions that would be most likely to have decodable internal representations. One recent study, for example, found that a remembered face was best decoded from temporal voxels (VanRullen & Reddy, 2019). For the same reason, we also focused on an experimental setup with a training set of faces from only one fMRI session. This led to a training set substantially smaller than other recent studies (Dado et al., 2022; Güçlütürk et al., 2017; VanRullen & Reddy, 2019). We designed a procedure that attempted to balance maximizing the total number of unique faces viewed with not overtaxing the attention of participants. There is also a tradeoff in the design between including more fast trials or slower trials that individually have better estimates. Our balancing of these tradeoffs led to a total of 396 unique training stimuli and 36 test images.

Ultimately, the most potentially applicable approach would involve not only limiting the training session to one scan session, but actually limiting it to only a portion of the session. In order to pursue this possibility, one important aspect of the present work is an investigation into the utility of across-participant reconstructions as opposed to the more typical within-participant approach. This approach has recently shown to improve the reconstruction of natural images (Akamatsu et al., 2021). Here, one important aspect of the design was the inclusion of two “fixed” runs where all participants saw the same images in the same order. These fixed runs make the data more amenable to transforming the functional data into a shared space across all participants (Chen et al., 2015; Chen et al., 2017). With this approach, the training set for each participant can be multiplied by the number of participants included. When applied to the present data, it has the potential to greatly improve the reconstruction accuracy. If that approach proves viable, it opens up the possibility of participants only needing to be shown those two fixed runs (or possibly one) in order to be transformed into this shared space. This would allow the rest of a scan session to be devoted to a specific experimental design that decodes a participant’s internal representation of face images over the course of the experiment.

Going forward, there are a number of additional analyses that are being actively explored that could potentially increase the power of this approach: (1) Use a leave one out analysis that ignores the test faces as distinct. This approach would make the measurement less noisy and less influenced by any particular test image. (2) Include a voxel selection technique, such as the inclusion of only face-selective voxels. This could help reduce the likelihood of overfitting the model. (3) Adapt this data to more recent approaches to face parameterization. This has greatly improved reconstructions in other contexts. (4) Leverage the relationship between particular regions and components to create reconstructions pieced together from the most reliably decoded region/component pairings. Although this would be less informative of any particular region, this could lead to better overall predictions. With these possibilities (and others) there is reason to believe the power of this approach can be improved.

Chapter V

GENERAL DISCUSSION

The goal of this dissertation was to establish an approach to studying interference resolution in episodic memory. Long-term memories are too often studied without full appreciation of their multi-dimensional and reconstructive nature. Further, cognitive and neural perspectives on this phenomenon often proceed on separate tracks. Thus, I sought to develop an approach that could bridge findings from behavioral and neural paradigms, and computational models of interference resolution. This integrated approach will be highly valuable in understanding how the human memory system is able to efficiently store so many potentially confusable memories. In establishing this approach, I have already gained some key insights into the mechanisms of interference resolution.

Integrated summary of results

I began (Chapter 2) by collecting and establishing the validity and reliability of several metrics describing a large face stimulus corpus. Specifically, I first landmarked a large sample of faces by hand. The 62 positions were landmarked with high reliability across raters and allowed me to implement an active appearance model (AAM). The AAM components represent a face space that can be used to generate and manipulate synthetic face stimuli. In Chapter 4, I established the utility of these components in reconstructing faces from evoked patterns of fMRI activity.

I next collected data on the sorting of faces based on similarity. I found that sorters were consistent in their grouping of faces. In Chapter 3 I describe an application that combines the sorting data with the AAM components. This approach involved generating eight distinct face “families” based on which faces tended to be grouped together. From this starting point, I manipulated the similarity of faces within an orthogonal face space where faces generated from the same family caused high interference, but faces from different families did not.

I also collected subjective ratings on all face images. I found that the ratings were reliable both within and across participants. In Chapter 3 I describe an application combining the AAM components with the ratings. I focused on mapping the AAM components to the two most reliably rated dimensions (affect, gender). However, other subjective dimensions could be utilized moving forward (attractiveness, trustworthiness, dominance). Critically, I found that participants were highly accurate in retrieving both features (affect, gender) from memory.

Beyond the validity of this approach to research, this technology helped to unlock key theoretical insights. In Chapter 3 I found compelling evidence that resolving memory interference is associated with systematic, subtle changes in how memory features are recalled. The key to establishing this finding was the ability to independently manipulate two perceptually-important dimensions. I was able to manipulate the faces to be similar enough to cause interference, but distinct enough (on one diagnostic dimension) to learn. By probing memory on the same dimensions the faces were manipulated on, I was able to measure feature memory in a continuous, perceptually-important space. I found that competition induced *repulsion*, where memories shifted away from their competitor on the diagnostic compared to the non-diagnostic dimension. I also found an increase in *precision*, again on the diagnostic compared to the non-diagnostic dimension. Both of these changes in recalled feature information were associated with better associative memory—suggesting that they play an adaptive role in interference resolution.

In Chapter 4 I began the process of mapping the face dimensions I developed to patterns of neural activity. The goal was to establish a method that could be used to measure potential changes in the neural representation of faces during competitive learning. First, I established the ability to reliably reconstruct face images by decoding AAM components from patterns of fMRI activity. I proceeded to explore which dimensions were best reconstructed and from which brain regions. I found evidence that face images were, overall, best reconstructed from occipital cortex. I also found a pattern where temporal and posterior parietal regions reconstructed shape components better than appearance components. I further identified specific AAM components and subjective dimensions that were the most strongly predicted. In particular, affect, gender, and trustworthiness were well predicted from the occipital region. Although the ability to generate image reconstructions is one important way to visualize the power of the model, ultimately the predictions made at the specific component or dimension level are the key to utilizing this approach experimentally (see Future directions, below).

Together, these results establish a role for both repulsion and increased precision in interference resolution. They also establish a set of methods that enable a path forward to linking these findings to adaptive neural changes also associated with interference resolution. Below I discuss the full theoretical implications and potential next steps.

The role of selective attention in interference resolution

The dominant theoretical focus of this dissertation is on a mechanistic account of interference resolution where learning through interleaved practice leads to changes in how feature-level information is remembered (repulsion, precision). It is important to consider how or whether these changes may be explained by behavioral strategies adopted by participants. Namely, given the task demands, participants were likely to actively develop strategies to learn the cue-face associations. One advantageous strategy to learning would be to attempt to identify the differences between competitive faces and then to selectively attend to those differences. In fact, a strategy of attending to differences has previously been shown to eliminate retrieval-induced forgetting (Smith & Hunt, 2000). Moreover, recent work has demonstrated that diagnostic features of competing memories are more strongly represented in in neural activity patterns during memory retrieval (Zhao et al., 2021).

While selective attention to diagnostic features potentially played a role in learning the competing associations, a selective attention account does not readily explain the key results of repulsion and precision found in Chapter 3. That is, an account based on increased attention to diagnostic features does not predict that memory for these features will fundamentally shift with learning. In contrast, our results indicate a systematic distortion in feature memory that occurs, with learning, for specific features of specific items. While attention can create perceptual distortions in certain instances, these distortions operate on the dimensions themselves. For example, during perceptual learning, repeatedly attending to a particular dimension may "stretch" the perception of that dimension (Goldstone, 1998; Nosofsky, 1986). Critically, this type of stretching would lead to increased precision for any item that could be perceived along that dimension—in other words, the stretching should generalize across items. However, an important design feature of my experiments is that I counterbalanced, across items, which feature (affect, gender) corresponded to the diagnostic versus non-diagnostic dimensions. Therefore, the increased precision I observed for competitive items cannot be explained by a global change (stretching) in the perception of affect or gender. Instead, the increases in precision were specific to individual items. Because these changes at the individual item level were predictive of a reduction in interference, I believe that these changes were a driving force in resolving interference.

While the role of selective attention to diagnostic dimensions is likely to play some role in interference resolution and is worth further investigation, the key point is that it fails to account for the findings of repulsion and precision that are the focus of this dissertation. Below I discuss an alternative theoretical framework that more readily accounts for these feature-level changes.

Relationship between behavioral findings and neural accounts of interference resolution

The hippocampus has important properties that reduce the likelihood of interference. Pattern separation helps create a unique neural code in CA3 when a new event is being encoded (Yassa & Stark, 2011). A key facet of pattern separation is that regardless of the constituent features of the event, the associated neural representation is orthogonalized. This reduces the likelihood that any two events will interfere with one another during memory retrieval. However, when two events are highly similar (e.g. two egg recipe videos), this automatic mechanism may not be able to prevent interference. In those instances, recent fMRI evidence indicates that there is an *experience-dependent* process in the hippocampus, *repulsion*, that resolves interference over the course of learning. This process shifts the neural representations of items with high similarity (Chanales et al., 2017) or that have caused more interference (Wanjia et al., 2021) to become more distinct. Hippocampal repulsion predicts interference reduction, suggesting that these neural changes may be playing a mechanistic role in resolving interference (Favila et al., 2016; Wanjia et al., 2021).

This pattern of *neural* repulsion mirrors our findings of *behavioral* repulsion and precision. Namely, our findings show that highly similar items shift to be recalled as more distinct, just as neural representations shift to become more distinct in similar experimental contexts. Both the neural and behaviorally-measured changes have been linked to interference resolution. However, neural repulsion has not yet been shown to be associated with changes in the recollection of memory features. One theory that could explain that connection comes from Hullbert and Norman (2015). They model neural repulsion as a process where the diagnostic features of competing items become strengthened and the non-diagnostic features become weakened (see Interference resolution, Chapter 1). From this perspective, these changes to memory feature representations are what underlie neural repulsion. If the neural representation of specific memory features become strengthened or weakened, it follows that those same features would be recalled differently.

A challenge I addressed is translating the Hullbert and Norman (2015) model predictions into specific memory measures; I found evidence for two potential accounts. The most straightforward is developing a more precise memory for the strengthened compared to the weakened feature (i.e. diagnostic compared to non-diagnostic dimension). Regardless of the accuracy, if there is a strong representation for a particular feature, we would expect repeated retrieval attempts to have lower variance. Alternatively (or additionally), when a diagnostic feature becomes a larger part of the overall memory representation, that “over-representation” may create an exaggerated difference compared to the competitive item (specifically on that feature). Both of these explanations represent adaptive distortions compatible with this computational model and more broadly with how the reconstructive nature of memory can facilitate our interactions with the world (Schacter et al., 2011).

Future directions

Importantly, although our behavioral findings fit with the Hullbert and Norman (2015) model of neural repulsion, the present data is insufficient to make a full causal connection. An important future goal is to investigate the extent to which behavioral and neural repulsion are related. Future experiments could track changes in memory features simultaneously through behavioral probes and trial-level neural decoding, both mapped to the same underlying face features. If feature memory shifts can be linked to shifts in decoding predictions, in a time-locked way, this would be strong evidence for a relationship between the two.

Although our approach in Chapter 4 framed the results in terms of reconstructing face images, the most important objective for future applicability is the ability to predict specific face dimensions. The results in Chapter 4 suggest that affect and gender are the best predicted subjective dimensions. These same dimensions were the focus of Chapter 3, making them strong candidates to focus on moving forward.

Although those dimensions are the most promising, there are still challenges that need to be overcome prior to implementing this approach. One of the most pressing challenges is the development of a brief fMRI training session that can reliably predict those dimensions. Optimally, the training session would be short enough to still allow time for additional scan runs focused on interference manipulations. The fMRI study in Chapter 4 included over an hour of functional scanning and 396 unique training images. Given the number of features used (thousands of voxels), the model is already data-starved, so reducing the number of trials without any other changes is not a viable approach. The most straightforward approach, would be the transformation of each participant's neural data into a shared feature space (Chen et al., 2015; Chen et al., 2017; Haxby et al., 2011) so that model training could be performed across participants. This would vastly expand the size of training set by allowing all trials across participants to be used for model training, with one participant iteratively held-out for model testing. This type of training set would continue to expand as more participants are added, making this approach potentially more powerful over time, especially if used over multiple, successive experiments. The key to utilizing this type of approach is the inclusion of a fixed scan session where every participant is presented with the same stimuli in the same sequence. This fixed scan session can then be used to "functionally align" data across participants. If this approach proves viable, future experiments would only need to include the fixed scan run in order to map a new participant's data into the shared space. This would allow the rest of the scan session to be devoted to an experiment that the model makes predictions on, rather than devoting more scan time to adding to the training set. This type of approach has proved successful in decoding natural images (Akamatsu et al., 2021), however it has not yet been applied to decoding face dimensions. It is unknown whether the success will translate and how long of a fixed scan run would be needed for this approach to work for faces.

An alternative approach could focus on developing a training session specifically designed to decode the target dimensions. By focusing on as few as two dimensions, rather than the full multi-dimensional face space, there are potentially more powerful design options. Chapter 4 utilized an event-related design, where each face stimulus was treated as an independent event. I viewed this as necessary to target all unique properties of each face. However, under a more targeted approach, one could utilize a blocked design where faces matched on the target dimensions are displayed in succession. An approach like that would leave out a lot of variability in faces that the current design detected, but it may be able to achieve a stronger detected signal for the dimensions of interest. This type of targeted approach could lead to high decodability in a more limited amount of time.

Neither of these potential approaches directly address another outstanding challenge: whether the ability to reconstruct from perception will translate to memory retrieval. Previous face reconstruction attempts have found the ability to reconstruct remembered faces at above chance levels in temporal cortex (VanRullen & Reddy, 2019) and posterior parietal cortex (Lee & Kuhl, 2016). Thus, one potential path to applying the decoding of face features to memory is to simply focus on specific regions where there is an established ability to reconstruct. Reconstruction accuracies, however, were lower compared to perception-based reconstructions, so this approach may not lead to decoding accuracies sufficiently high to track changes in the representation between trials.

One alternative approach would be to train models on memory. Training on long-term memory would likely lead to too small of a training set (due to practical limits in how many faces a participant would be able to memorize in a single experimental session), however, training on working memory could be viable (see Lee & Kuhl, 2016). Another possibility would involve learning a more complex mapping to translate activation patterns in perception to memory. Although we know that during retrieval, patterns elicited during encoding are reinstated, there is a lot of unaccounted for variance in reinstatement patterns (Xue, 2018). There are a number of potential explanations (e.g. a similar representation that is merely depressed, an alteration in how features are represented, a bias towards higher-level features, or a shift in where items are primarily processed), all of which may play a role to different extents depending on the brain region (Favila et al., 2022). Accounting for these differences could be key to unlocking the full power of this method.

Broader implications

I established evidence that memories are systematically altered at the specific feature level by the presence of other memories. This has important implications for the measure of memory generally, particularly in the context of the growing use of continuous memory measures. Given the present findings, there are several important considerations for memory researchers: (1) Changes in one memory feature do not necessarily mean changes on other memory features. I found changes on the diagnostic *relative* to the non-diagnostic dimension. This is taken into account when it is the target of experimental manipulation, but often isn't when one probed feature is intended to measure overall memory accuracy, bias, or precision. (2) Different dimensions may demonstrate unique memory properties. I found initial evidence for differences in memory between affect and gender (see Fig. S3.1). Although these dimensions were counter-balanced and the differences did not impact our results of interests, these factors can lead to overall distortions in memory (Bülthoff & Zhao, 2019; Won et al., 2020). Again, the impact of this can be missed if only one feature is probed. (3) Adaptive (or maladaptive) feature memory changes must be taken into account. I found a systematic pattern of repulsion, a study that views memory only in reference to the veridical value would view that type of adaptive distortion as inaccurate. (4) Precision should not be conflated with accuracy. I found a pattern where repulsed memories were also highly precise. (5) Unique, item-level memory properties should be taken into account. The pattern of repulsion I found varied across items. Examining our results this way helped find evidence that greater repulsion was associated with reduced interference. Failure to take into account differences in feature memory bias between items could also lead to deflated estimates of precision, if each item is highly precise, but has a distinct degree of bias.

Outside of memory, these tools and approaches offer a path to innovation in a variety of domains. This dissertation focused on face stimuli because of how adept humans are at processing and remembering them, and due to their ability to be parameterized along perceptually-important, continuous dimensions. These properties make face stimuli useful in a wide variety of research applications. Although this dissertation focused on separate behavioral and neural designs, there is a clear path forward for the pursuit of convergent designs. When there are changes in a neural representation, there are likely behavioral consequences. It is the job of cognitive neuroscientists to identify those changes, even in more abstract domains like episodic memory (Krakauer et al., 2017). Inspired by cognitive models and a variety of neural findings, this dissertation identified and translated models of neural computations into specific, behavioral consequences. The tools validated here offer further opportunities for the translation of experimental manipulations along comparable, but behavioral- or neural-derived measures.

Conclusion: how do we remember highly similar information?

Let’s return to the question I began this dissertation with: how do humans store so many memories? First, not every memory needs to be remembered as distinct. Some memories are best forgotten; others are best integrated into semantic knowledge or schemas without a full episodic memory specific to the event. When it is advantageous to form a distinct memory, there is a dedicated path of the memory system that helps avoid interference. This system, however, is not always effective and memories can be (and are quite often) forgotten due to other memories. However, in cases where there is additional pressure on the system to store a once forgotten or interference-prone memory, interference can be overcome. The exact neural computation involved in this process are still being studied. Here, I established two adaptive memory changes that help reduce interference within the time course of the experiment. These feature memory changes highlight or even exaggerate differences between competing memories in a way that could prevent forgetting over the long-term. I propose ways this could relate to underlying neural processes that repulse the representation of competing memories and chart a path forward to establishing the full relationship between the two. Such an interference-resolution mechanism would explain the vast human memory capacity.

APPENDIX A

CHAPTER II SUPPLEMENTARY MATERIAL

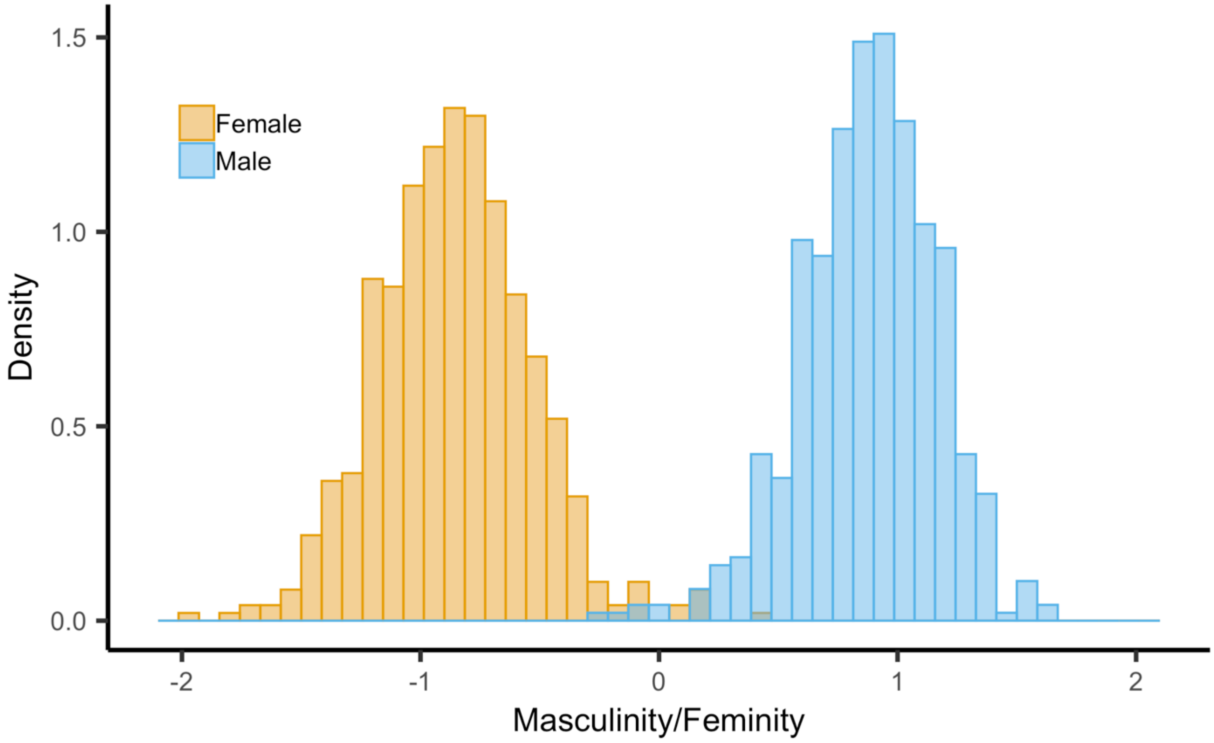

Figure S2.1. Density plot of masculinity/femininity ratings across all stimuli, divided by hand labels of female (orange) and male (blue) perceived gender.

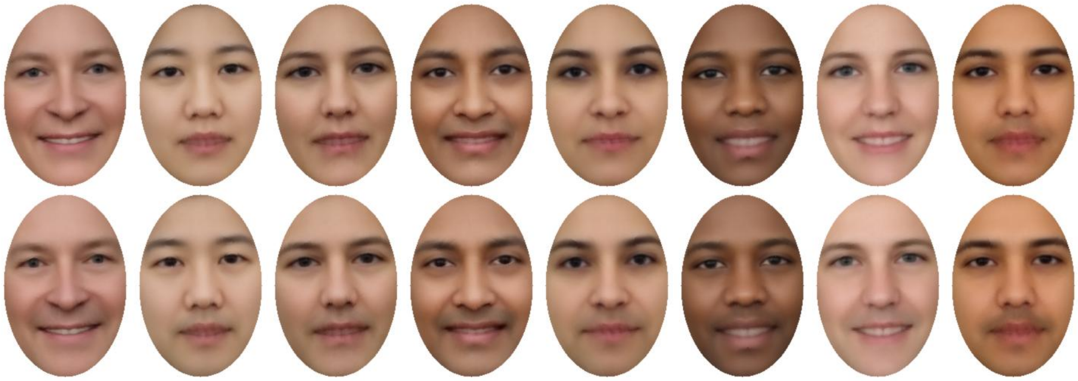

Figure S2.2. Example of 8 stimulus pairs (left to right) from Chapter 3 (experiment 1) matched on affect, but differing on gender (top to bottom).

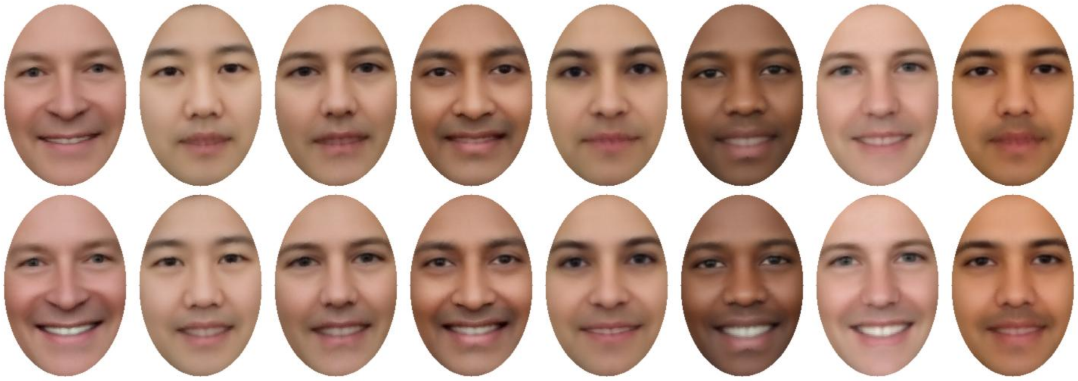

Figure S2.3. Example of 8 stimulus pairs (left to right) from Chapter 3 (experiment 1) matched on gender, but differing on affect (top to bottom).

APPENDIX B

CHAPTER III SUPPLEMENTARY MATERIAL

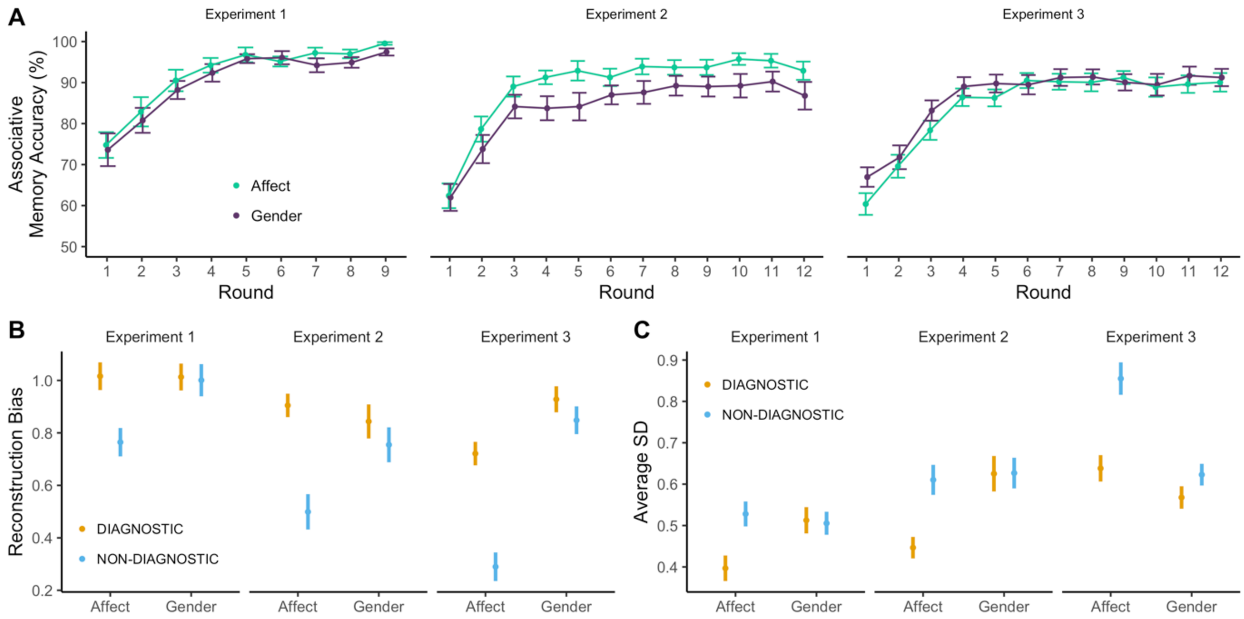

Figure S3.1. Differential memory effects for affect vs. gender. **A.** Accuracy (percent correct) on the associative memory test during each round of the learning phase (competitive condition only) separated by the whether the diagnostic dimension was affect (green) or gender (purple) and by experiment number. For exp. 1, there was no difference in accuracy for affect vs. gender ($F(1,35) = 1.84$, $p = 0.18$, $\eta_G^2 = 0.004$). For exp. 2, the similarity between competitive pairmates was increased on both affect and gender (see Methods). Although not intended, this resulted in accuracy being significantly higher when affect was the diagnostic dimension compared to gender ($F(1,40) = 13.22$, $p < 0.001$, $\eta_G^2 = 0.026$). We addressed this difference in exp. 3 by further (and selectively) increasing the similarity between competitive pairmates along the affect dimension. This change was successful, as there was no longer a significant difference in accuracy when the diagnostic dimension was affect vs. gender ($F(1,56) = 2.22$, $p = 0.14$, $\eta_G^2 = 0.003$). **B.** Reconstruction bias as a function of dimension (diagnostic, non-diagnostic), whether the diagnostic dimension was affect or gender, and experiment number Here, bias was measured as the (un-modeled) mean response because there were too few trials to perform the modelling approach used in the main text. A repeated measures ANOVA revealed a robust main effect of dimension, reflecting significantly greater bias towards repulsion (higher mean reconstruction bias) on the diagnostic vs. non-diagnostic dimension ($F(1,118) = 85.05$, $p < 0.001$, $\eta_G^2 = 0.089$). However, there was also a significant interaction between diagnostic/non-diagnostic dimension and gender/affect ($F(1,118) = 34.41$, $p < 0.001$, $\eta_G^2 = 0.047$) reflecting a greater difference between diagnostic vs. non-diagnostic dimensions when the diagnostic dimension was affect (this effect did not further interact with experiment number, three-way interaction: $F(2,118) = 0.39$, $p = 0.68$, $\eta_G^2 = 0.001$). Nonetheless, the effect of diagnostic vs. non-diagnostic dimension was significant when the diagnostic dimension was affect ($F(1,118) = 88.44$, $p < 0.001$, $\eta_G^2 = 0.234$) or gender ($F(1,118) = 4.27$, $p = 0.041$, $\eta_G^2 = 0.008$). **C.** Reconstruction precision (SD of responses across the 4 reconstruction trials for each face) as a function of dimension (diagnostic, non-diagnostic), whether the diagnostic dimension was affect or gender, and experiment number. A repeated measures ANOVA revealed a robust main effect of dimension ($F(1,118) = 45.30$, $p < 0.001$, $\eta_G^2 = 0.054$), reflecting greater precision (lower SD) for the diagnostic than the non-diagnostic dimension. However, there was also a significant interaction between diagnostic/non-diagnostic dimension and gender/affect ($F(1,118) = 36.60$, $p < 0.001$, $\eta_G^2 = 0.034$), reflecting a greater difference between the diagnostic vs. non-diagnostic dimensions when the diagnostic dimension was affect (this effect did not further interact with experiment number, three-way interaction: $F(2,118) = 0.086$, $p = 0.92$, $\eta_G^2 < 0.001$). When affect was the diagnostic dimension, the effect of diagnostic vs. non-diagnostic dimension was significant ($F(1,118) = 76.34$, $p < 0.001$, $\eta_G^2 = 0.147$). In contrast, when gender was the diagnostic dimension, the effect of diagnostic vs. non-diagnostic dimension was not significant ($F(1,118) = 1.22$, $p = 0.27$, $\eta_G^2 = 0.003$). Note: error bars represent SEM.

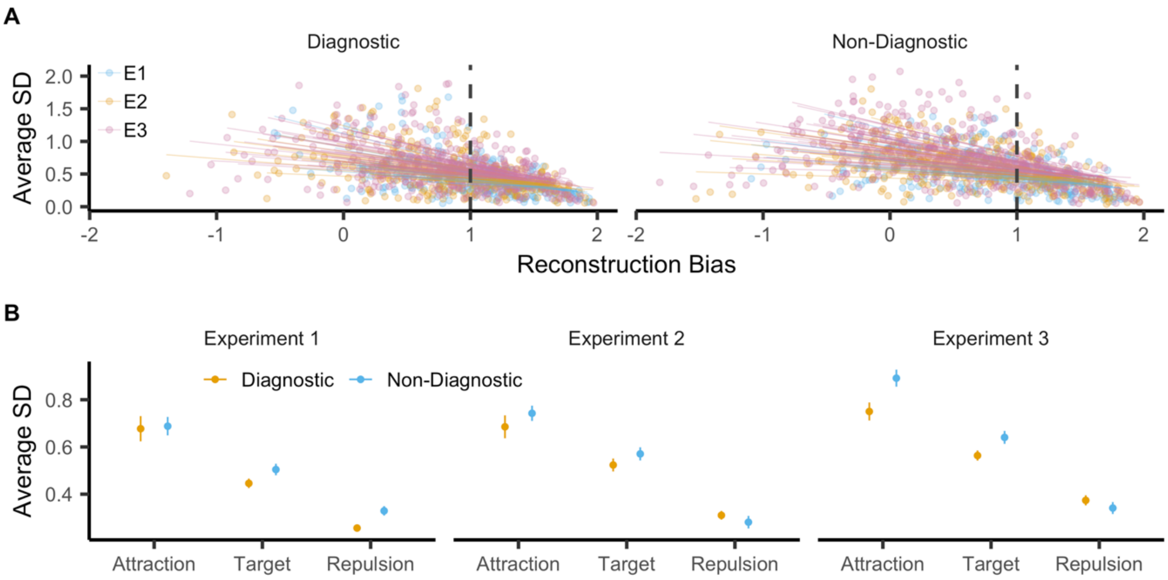

Figure S3.2. Relationship between reconstruction bias and precision. **A.** A mixed-effects model was run with precision (mean SD) as the dependent variable and with experiment number (1, 2, 3), dimension (diagnostic, non-diagnostic) and bias included as predictors. Measures of precision and bias were computed at the item level, with each measure based on the mean value across the 4 reconstruction trials for each face. The relationship between bias and precision was modeled with random intercepts and slopes for each participant. Compared against a model without bias, adding bias significantly improved model fit ($\chi^2(1) = 140.8$, $p < 0.001$, $\beta_{bias} = -0.25$, $SE = 0.015$) reflecting the fact that stronger bias (repulsion) was associated with greater precision (lower SD). In the plot, each dot represents a specific face image, each experiment is a unique color (e1: blue; e2: orange; e3: pink), and each line represents the modelled, participant-specific relationship between reconstruction bias and precision. Notably, the effect of bias on precision did not interact with dimension type (diagnostic vs. non-diagnostic: $\chi^2(1) = 1.61$, $p = 0.20$, $\beta_{biasXdim} = 0.027$, $SE = 0.022$, or experiment: $\chi^2(2) = 2.32$, $p = 0.31$). Thus, although the diagnostic dimension was associated with greater bias and precision compared to the non-diagnostic dimension (see main text), the relationship between bias and precision was not specific to the diagnostic dimension. **B.** One potential account of why precision was greater on the diagnostic dimension is that greater repulsion (towards the boundary) reduced the response space for reconstruction (compressing variance). To address this, we computed mean precision for each dimension as a function of the level of bias on the diagnostic and non-diagnostic dimensions. Three equal-width bias bins were created within the half of the response range that was closer to the target than the competitor (0-2): ‘Attraction’ represents bias in the direction of the competitor face (range of bias values: 0-0.67); ‘Target’ represents responses centered around the true value (range: 0.67-1.33); ‘Repulsion’ represents bias away from the competitor face (range: 1.33-2). Because not all participants contributed to each bin, the mean and SEM were calculated across items, ignoring participant. Qualitatively, while precision was markedly higher, overall, for the Repulsion bin (lower SD), the tendency for greater precision on the diagnostic vs. non-diagnostic dimension was *not* selective to the Repulsion bin—in fact, the effect was least consistent in the Repulsion bin. In order to statistically confirm that the difference in precision for diagnostic vs. non-diagnostic dimensions was not an artifact of high bias values, we calculated the mean precision for the diagnostic and non-diagnostic dimensions only including faces within the Attraction and Target bins (0–1.33). Data were included from all experiments, but one participant (from e3) was excluded for having no items within the specified range. The remaining participants each had at least 2 items in the specified range for both the diagnostic ($M = 5.70$) and non-diagnostic ($M = 5.16$) dimensions (out of 8 possible items). Even with this restricted range (that excluded high bias items), there was significantly greater precision on the diagnostic compared to the non-diagnostic dimension ($F(1,117) = 25.16$, $p < 0.001$, $\eta^2_G = 0.051$). This effect did not interact with experiment ($F(2,117) = 2.20$, $p = 0.12$, $\eta^2_G = 0.009$). Note: error bars represent SEM.

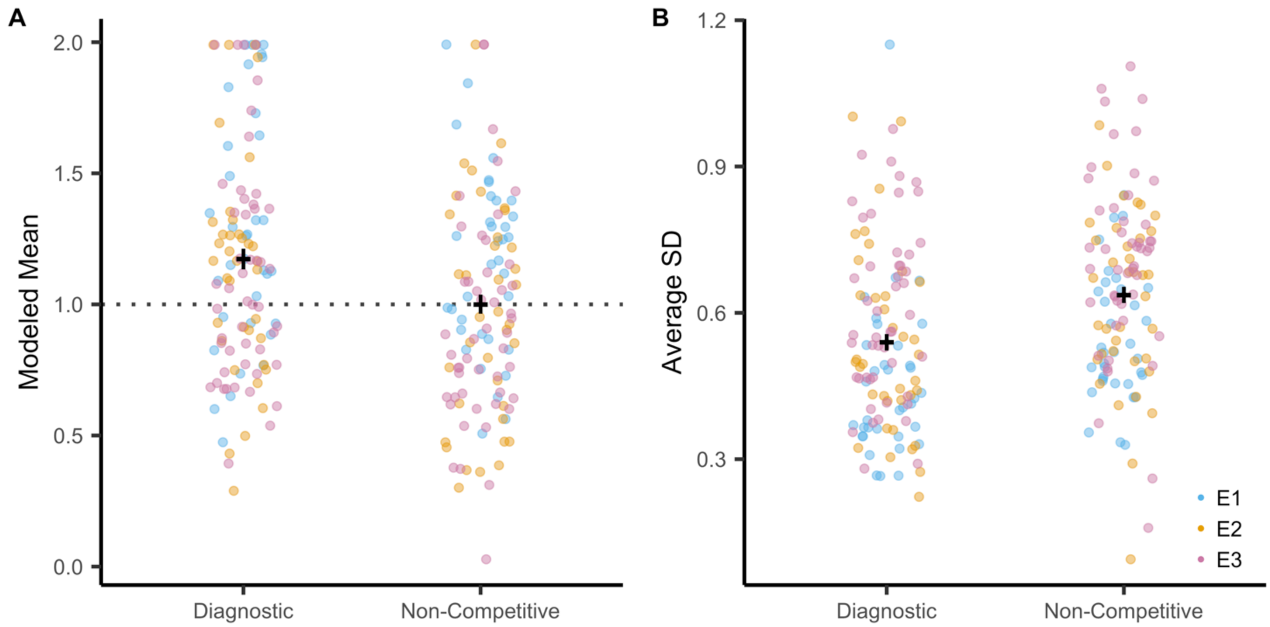

Figure S3.3. Reconstruction bias and precision for faces in the competitive vs. non-competitive conditions. For the non-competitive condition there was no distinction between diagnostic vs. non-diagnostic dimensions. Thus, for each face in the non-competitive condition, data from both dimensions were included. With 4 items in the non-competitive condition for each participant, and 4 reconstruction trials per face, this yielded 32 total values per participant for the non-competitive condition (2 dimensions x 4 items x 4 trials). Bias was modeled using the same method as for the diagnostic and non-diagnostic dimensions (see Methods). **A.** Bias was significantly greater (higher modeled mean) for the diagnostic dimension (of faces in the competitive condition) than for the non-competitive condition ($F(1,118) = 22.11$, $p < 0.001$, $\eta_G^2 = 0.043$). This difference did not interact with experiment ($F(2,118) = 0.13$, $p = 0.88$, $\eta_G^2 < 0.001$). There was also a significant difference between the non-diagnostic dimension and the non-competitive condition ($F(1,118) = 6.73$, $p = 0.011$, $\eta_G^2 = 0.015$; not shown in the figure), with no interaction by experiment ($F(2,118) = 1.90$, $p = 0.15$, $\eta_G^2 = 0.008$). Specifically, for the non-diagnostic dimension there was a relative bias toward the center of face space (modeled mean tending to be lower than 1; see **Figure 3.4**) whereas for the non-competitive condition the modeled mean was higher (almost exactly at the true value of 1). **B.** Precision was significantly greater (lower mean SD) for the diagnostic dimension (of faces in the competitive condition) than for the non-competitive condition ($F(1,118) = 39.44$, $p < 0.001$, $\eta_G^2 = 0.073$). This difference did not interact with experiment ($F(2,118) = 0.12$, $p = 0.89$, $\eta_G^2 < 0.001$). Notably, there was no significant difference in precision between the non-competitive condition and the non-diagnostic dimension ($F(1,118) = 0.006$, $p = 0.94$, $\eta_G^2 < 0.001$; not shown in the figure), nor was there an interaction by experiment ($F(2,118) = 1.75$, $p = 0.18$, $\eta_G^2 = 0.006$). Notes: Each dot represents a participant, with color indicating the experiment (e1: blue; e2: orange; e3: pink); error bars represent SEM.

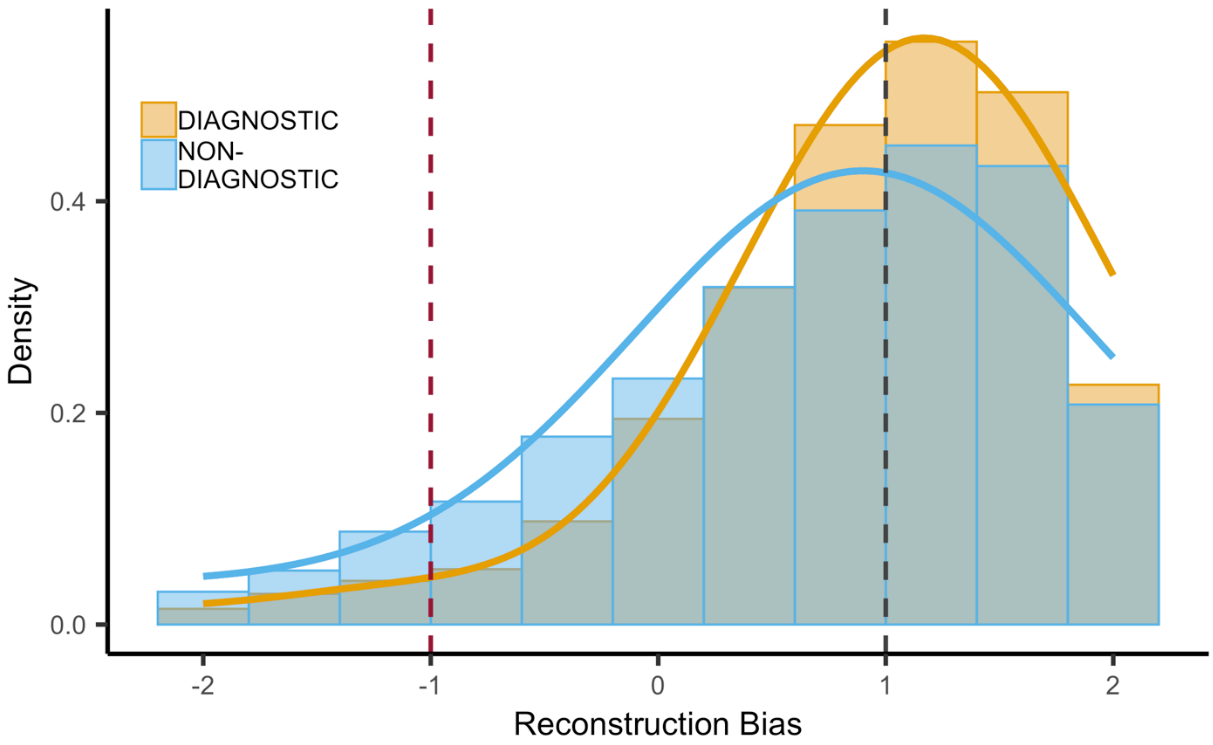

Figure S3.4. Histogram of reconstruction responses across all experiments, participants, and items in the competitive condition. Responses were separated by whether the dimension was diagnostic (orange) or non-diagnostic (blue). As in all other analyses, responses were rescaled such that the location of the target was at 1 (black dotted line), the center of the face space was at 0, and in the case of the diagnostic dimension, the location of the competitor was at -1 (red dotted line). To better characterize the distributions, separate mixture models were generated for the diagnostic and non-diagnostic dimensions. Each model included three distributions: a target distribution (the correct face), a competitor distribution (the competitor face), and a uniform distribution (random guessing). For the target distribution, we used a truncated normal distribution where we set the mean to the mean of our estimate from the main bias analysis across all participants from all experiments (diagnostic: 1.17; non-diagnostic: .9) and allowed the standard deviation to vary within a 'generous' range that was wide beyond what would plausibly explain the data (0.3–2). We used the same approach for the competitor distribution but changed the mean. For the diagnostic dimension, we mirrored the target bias value by setting the competitor value to -1.17. For the non-diagnostic dimension, since there was no competitor, we set the competitor value at the value where a competitor would be (-1). Although there was no competitor in the case of the non-diagnostic dimension, we included it here to allow a fairer comparison across the diagnostic and non-diagnostic dimensions. In particular, the non-diagnostic dimension allows for a baseline estimate of the percentage of swap errors (recalling the competitor) in a situation where there should not be any. For the diagnostic dimension, the best fitting model estimated that the target distribution explained 91.9% of responses (SD = .8), as reflected by the orange line. The model estimated that 6.1% of responses were random guesses and 2.0% of responses were swap errors (SD = .5). For the non-diagnostic dimension, the best fitting model estimated that the target distribution explained 84.1% of responses (SD = 1), as reflected by the blue line. The model estimated that 15.9% of responses were random guesses and 0% were swap errors. Taken together, these mixture model results suggest that the target distributions largely explained responses, with relatively little influence from random guesses and swap errors. That said, because the mixture models require a relatively high number of data points, these models were not well-suited to characterizing distributions for individual items (faces) and participants.

APPENDIX C

CHAPTER IV SUPPLEMENTARY MATERIAL

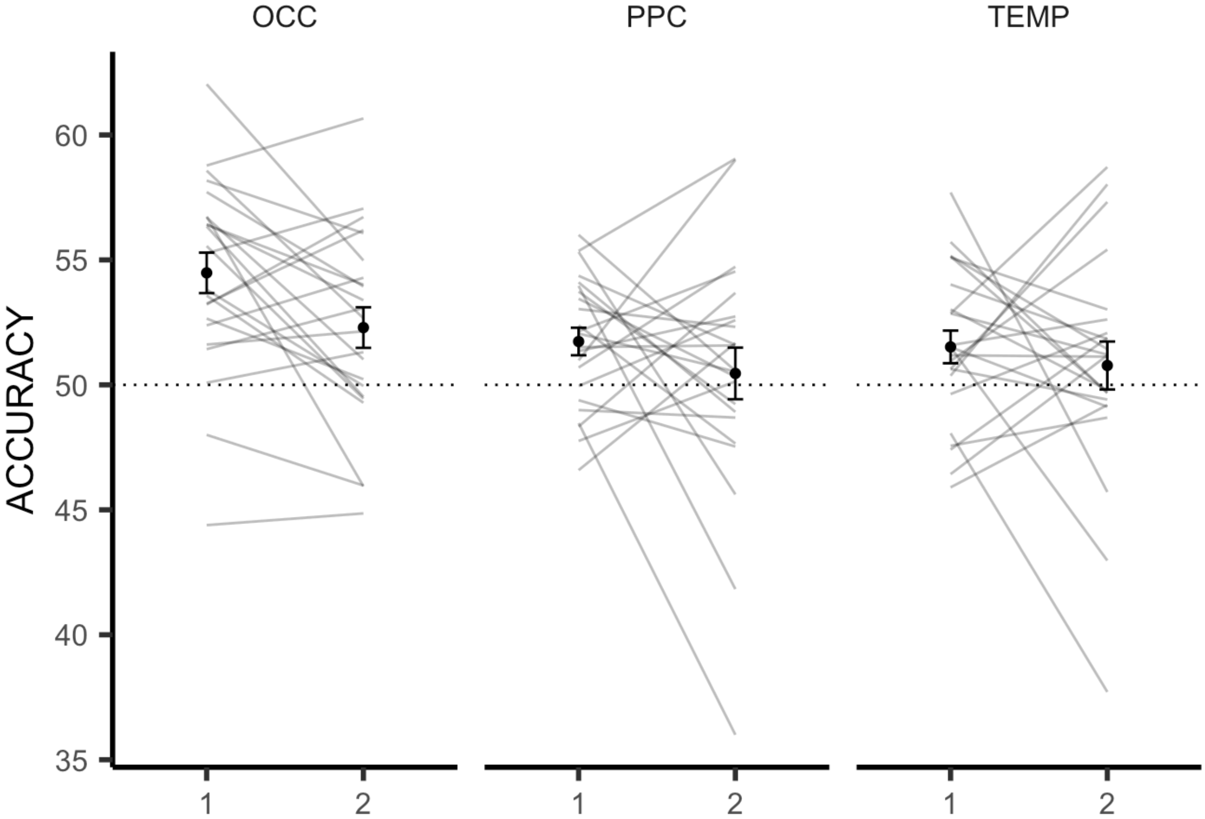

Figure S4.1. Alternative forced choice accuracy (AFC) for correct trials only, modeled separately for 1st and 2nd appearance (x-axis), and for occipital (OCC), posterior parietal (PPC), and temporal (TEMP) cortical ROIs. For ease of interpretation, the current plot and results focus on participants (N=23) who only saw the test items twice. A repetition (1, 2) x ROI (OCC, PPC, TEMP) repeated measures ANOVA found no significant difference in repetition (*F*(1, 22) = 2.45, *p* = 0.13, η_G^2 = 0.03) and no interaction (*F*(2, 44) = 1.4, *p* = 0.26, η_G^2 = 0.01). There was a significant effect of ROI (*F*(2, 44) = 11.54, *p* < 0.001, η_G^2 = 0.07). Follow up t-tests revealed a significant drop in performance between repetitions 1 and 2 in OCC (*t*(22) = 2.76, *p* = 0.01, *d* = 0.56), but not in PPC (*t*(22) = 1.13, *p* = 0.27, *d* = 0.32) or TEMP (*t*(22) = 0.66, *p* = 0.52, *d* = 0.19). Error bars represent SEM

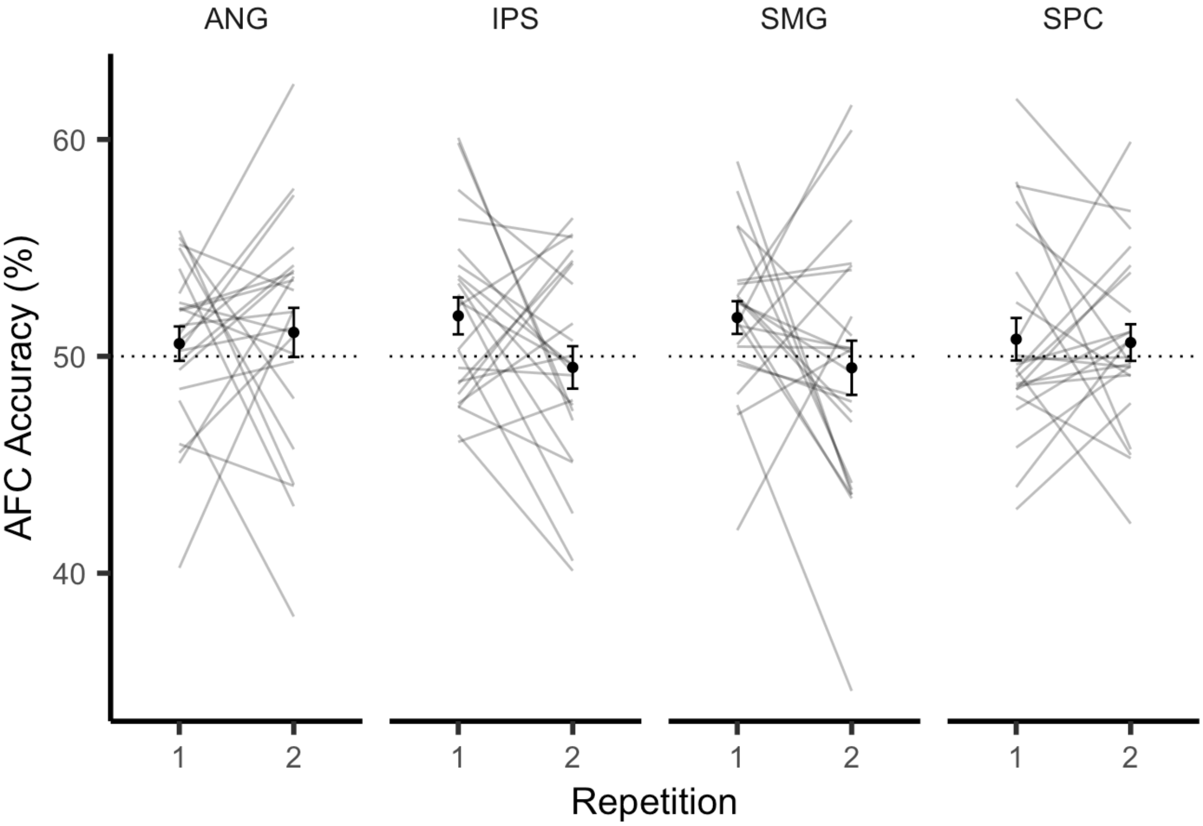

Figure S4.2. Alternative forced choice accuracy (AFC) for correct trials only, modeled separately for 1st and 2nd appearance (x-axis), and for ROIS within the overall PPC region: angular gyrus (ANG), intraparietal sulcus (IPS), supramarginal gyrus (SMG), and superior parietal cortex (SPC). For ease of interpretation, the current plot and results focus on participants (N=23) who only saw the test items twice. A repetition (1, 2) x ROI (ANG, IPS, SMG, SPC) repeated measures ANOVA found no significant difference in repetition (*F*(1, 22) = 1.95, *p* = 0.18, η_G^2 = 0.01) and no interaction (*F*(3, 66) = 1.35, *p* = 0.27, η_G^2 = 0.02). There was also no significant effect of ROI (*F*(3, 66) = 0.03, *p* = 0.99, η_G^2 < 0.01). Follow up t-tests revealed no significant differences in any ROI between repetition 1 and 2. Error bars represent SEM

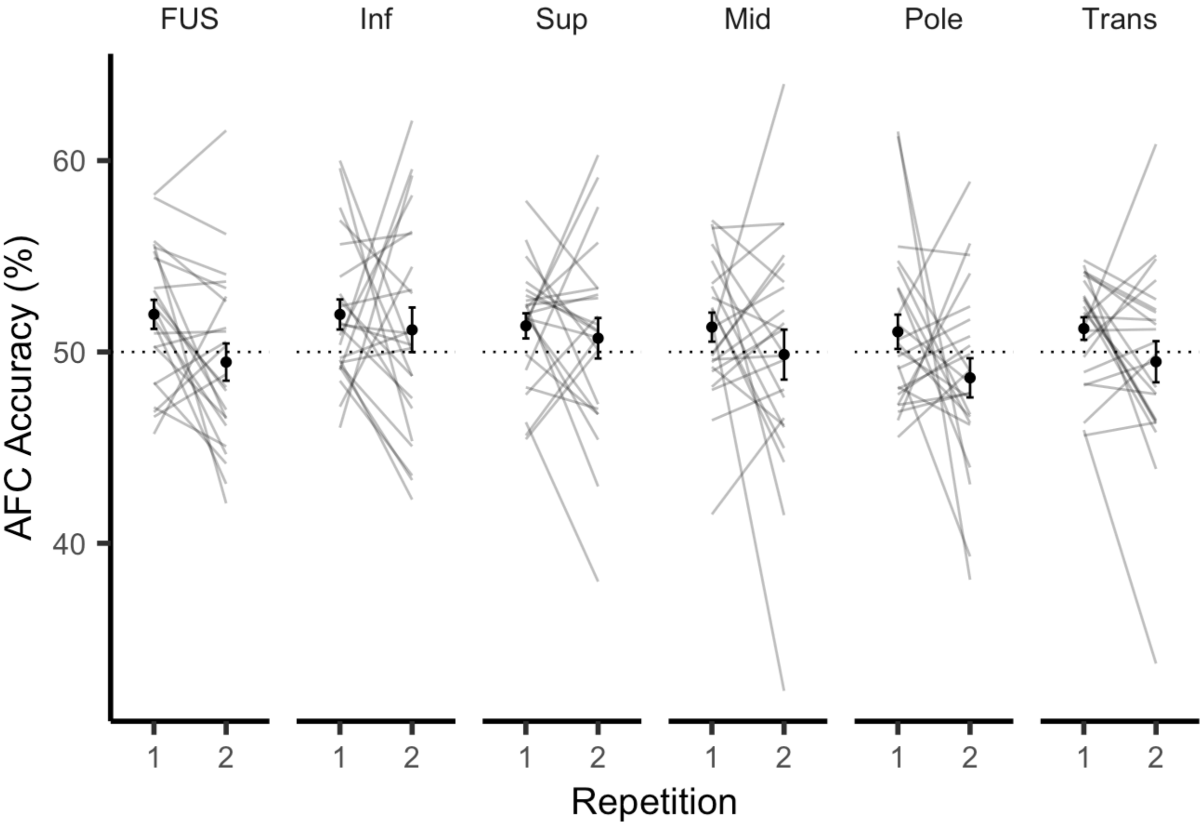

Figure S4.3. Alternative forced choice accuracy (AFC) for correct trials only, modeled separately for 1st and 2nd appearance (x-axis), and for ROIS within the overall temporal ROI: inferior temporal (Inf), superior temporal (Sup), middle temporal (Mid), temporal pole (Pole), and transverse temporal (Trans). We also included the fusiform gyrus (FUS), which was not included in the overall temporal ROI. For ease of interpretation, the current plot and results focus on participants (N=23) who only saw the test items twice. A repetition (1, 2) x ROI (Inf, Sup, Mid, Pole, Trans, FUS) repeated measures ANOVA found no significant difference in repetition (*F*(1, 22) = 2.93, *p* = 0.10, η_G^2 = 0.03) and no interaction (*F*(5, 110) = 0.57, *p*= 0.72, η_G^2 = 0.01). There was also no significant effect of ROI (*F*(5, 110) = 1.21, *p* = 0.31, η_G^2 = 0.01). Follow up t-tests revealed that only FUS significantly differed between repetition 1 and 2, with performance falling on the repeated trial (*t*(22) = 2.52, *p* = 0.02, *d* = 0.59). Error bars represent SEM

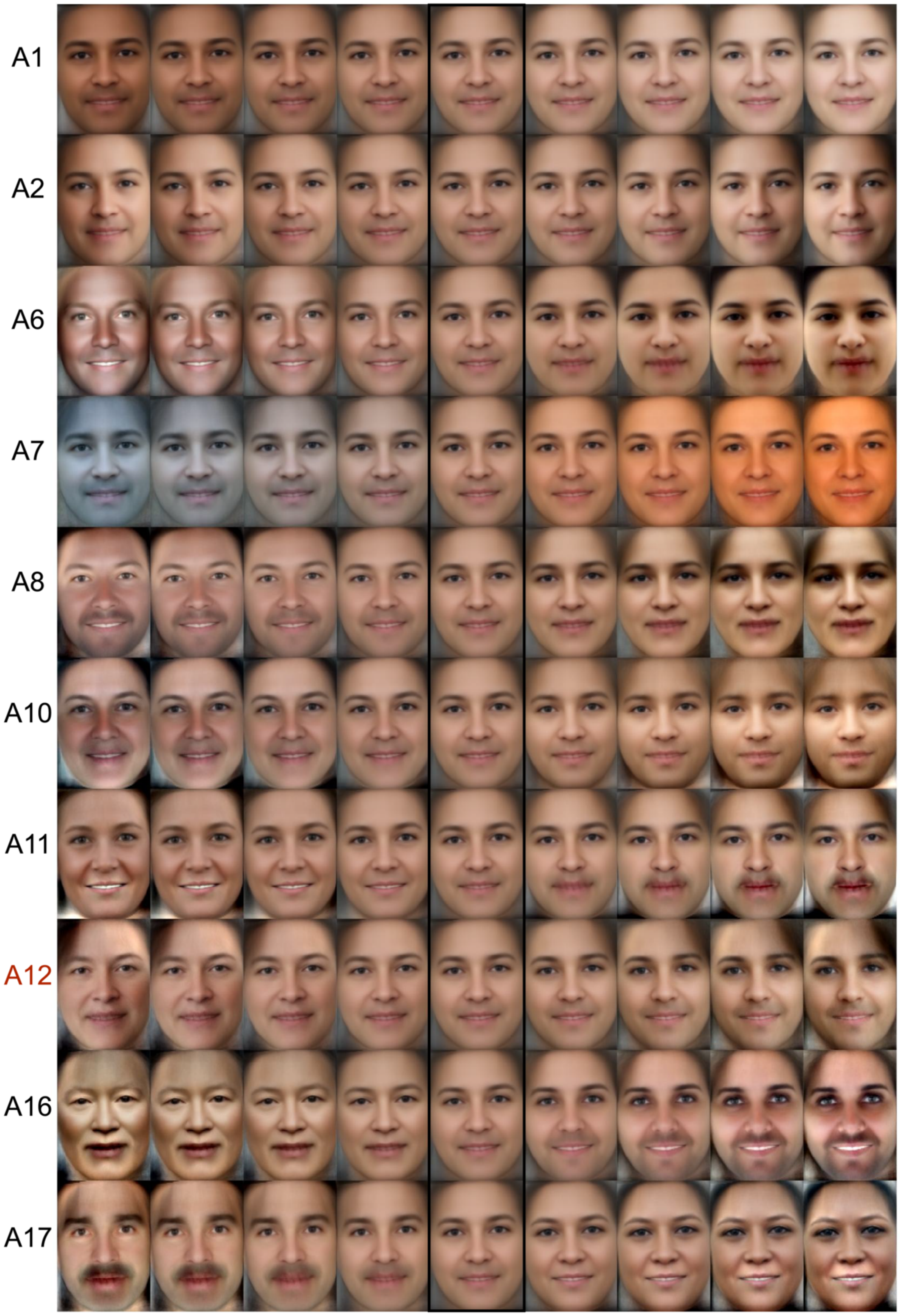

Figure S4.4. AAM appearance components significantly predicted by only OCC (black) and only PPC (red). The mean face (center) is depicted, shifted uniform amounts for each component (rows).

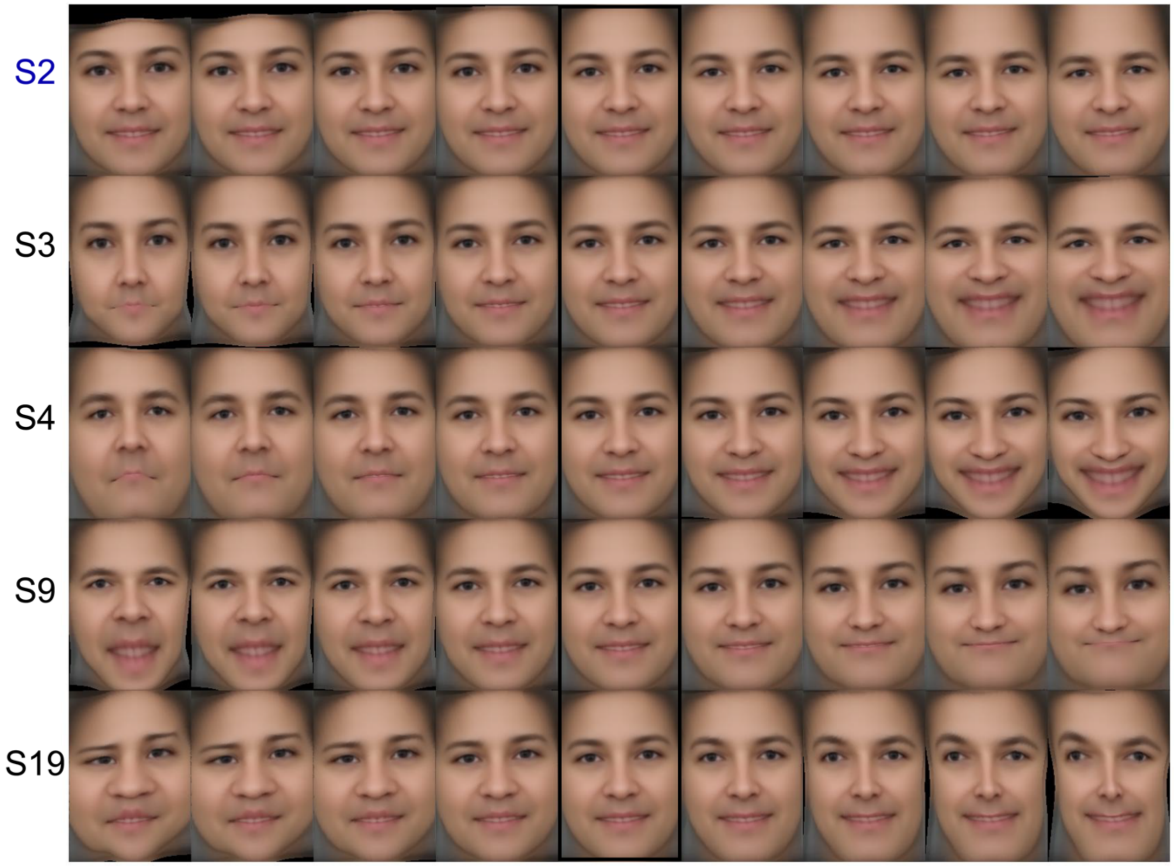

Figure S4.5. AAM shape components significantly predicted by only OCC (black) and by both OCC and TEMP (blue). The mean face (center) is depicted, shifted uniform amounts for each component (rows).

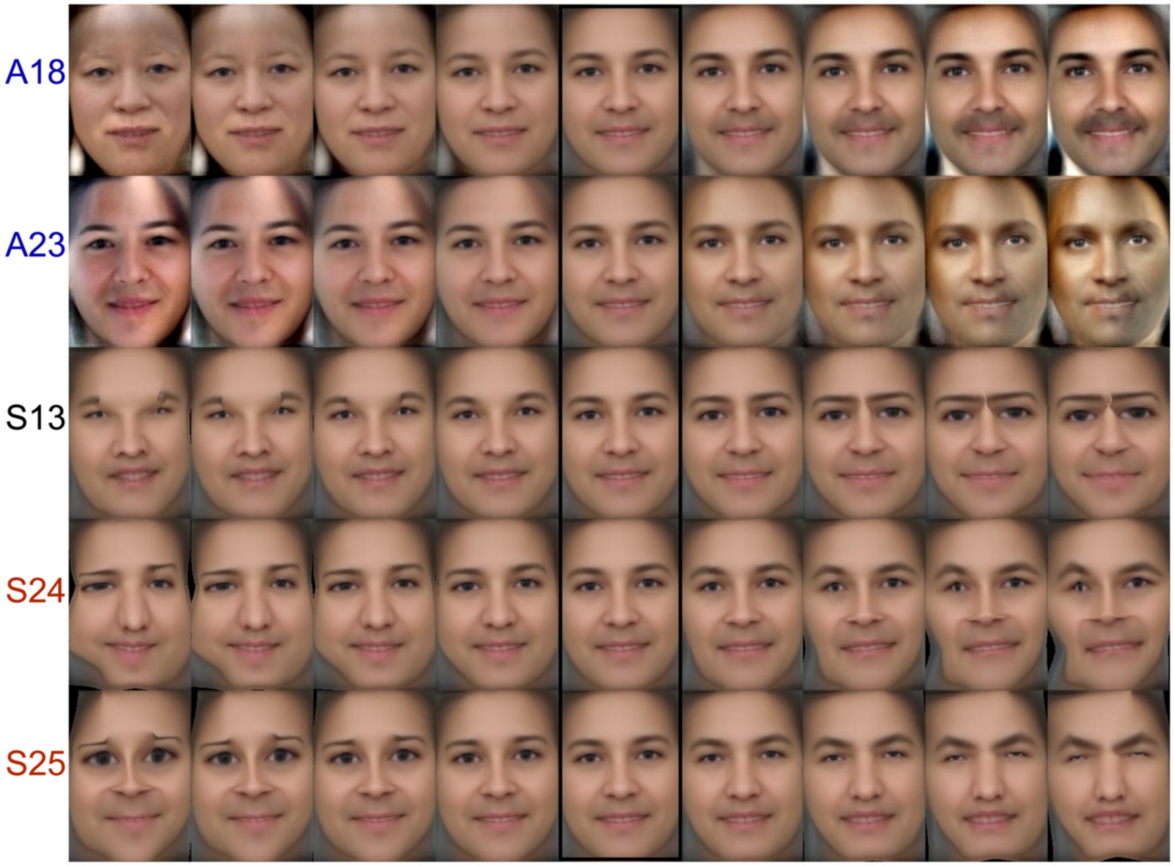

Figure S4.6. AAM components with significant negative correlations, for all ROIs (black), TEMP only (blue), and PPC only (red). The mean face (center) is depicted, shifted uniform amounts for each component (rows).

REFERENCES CITED

Akamatsu, Y., Harakawa, R., Ogawa, T., & Haseyama, M. (2021). Perceived image decoding from brain activity using shared information of multi-subject fMRI data. *IEEE access*, 9, 26593-26606.

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

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic bulletin & review*, *7*(3), 522-530.

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

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In *Memory* (pp. 237-313). Academic Press.

Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: memory retrieval as a model case. *Psychological review*, *102*(1), 68.

Archibald, F. (2009). Warping Using Thin Plate Splines. *MATLAB Central File Exchange*.

Arsalidou, M., Morris, D., & Taylor, M. J. (2011). Converging evidence for the advantage of dynamic facial expressions. *Brain topography*, 24(2), 149-163.

Ashby, S. R., Bowman, C. R., & Zeithamova, D. (2020). Perceived similarity ratings predict generalization success after traditional category learning and a new paired-associate learning task. *Psychonomic bulletin & review*, 27(4), 791-800.

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1), 26-41.

Baddeley, A. D. (1964). Semantic and acoustic similarity in short-term memory. *Nature*, 204(4963), 1116-1117.

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

Bae, G. Y., & Luck, S. J. (2017). Interactions between visual working memory representations. *Attention, Perception, & Psychophysics*, *79*(8), 2376-2395.

Bainbridge, W.A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face images. *Journal of Experimental Psychology: General*, 142(4), 1323-1334

Bakker, A., Kirwan, C. B., Miller, M., & Stark, C. E. (2008). Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science*, 319(5870), 1640-1642.

Balas, B., & Pacella, J. (2015). Artificial faces are harder to remember. *Computers in human behavior*, 52, 331-337.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3), 255-278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv.* https://doi.org/10.48550/arXiv.1406.5823

Battaglia, F. P., Benchenane, K., Sirota, A., Pennartz, C. M., & Wiener, S. I. (2011). The hippocampus: hub of brain network communication for memory. *Trends in cognitive sciences*, 15(7), 310-318.

Bäuml, K. H., & Hartinger, A. (2002). On the role of item similarity in retrieval-induced forgetting. *Memory*, 10(3), 215-224.

Bays, P. M., Catalao, R. F., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of vision*, 9(10), 7-7.

Beliy, R., Gaziv, G., Hoogi, A., Strappini, F., Golan, T., & Irani, M. (2019). From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI. *Advances in Neural Information Processing Systems*, 32.

Benda, M. S., & Scherf, K. S. (2020). The Complex Emotion Expression Database: A validated stimulus set of trained actors. *PloS one*, 15(2), e0228248.

Benoit, R. G., & Schacter, D. L. (2015). Specifying the core network supporting episodic simulation and episodic memory by activation likelihood estimation. *Neuropsychologia*, 75, 450-457.

Berens, S. C., Richards, B. A., & Horner, A. J. (2020). Dissociating memory accessibility and precision in forgetting. *Nature Human Behaviour*, *4*(8), 866-877.

Bjork, R. A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. *Varieties of memory and consciousness: Essays in honour of Endel Tulving*, 309-330.

Bonnici, H. M., Richter, F. R., Yazar, Y., & Simons, J. S. (2016). Multimodal feature integration in the angular gyrus during episodic and semantic retrieval. *Journal of Neuroscience*, 36(20), 5462-5471.

Bookstein, F. L. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6), 567-585.

Bostock, E., Muller, R. U., & Kubie, J. L. (1991). Experience-dependent modifications of hippocampal place cell firing. *Hippocampus*, 1(2), 193-205.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38), 14325-14329.

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological science*, *24*(6), 981-990.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision*, *10*(4), 433-436.

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience*, 29(44), 13992-14003.

Brunec, I. K., Moscovitch, M., & Barense, M. D. (2018). Boundaries shape cognitive representations of spaces and events. *Trends in Cognitive Sciences*, 22(7), 637-650.

Bülthoff, I., & Zhao, M. (2020). Personally familiar faces: Higher precision of memory for idiosyncratic than for categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(7), 1309.

Cao, R., Li, X., Todorov, A., & Wang, S. (2020). A flexible neural representation of faces in the human brain. *Cerebral Cortex Communications*, 1(1), tgaa055.

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of cognitive neuroscience*, 15(5), 704-717.

Chadwick, M. J., Hassabis, D., & Maguire, E. A. (2011). Decoding overlapping memories in the medial temporal lobes using high-resolution fMRI. *Learning & Memory*, 18(12), 742-746.

Chanales, A. J., Oza, A., Favila, S. E., & Kuhl, B. A. (2017). Overlap among spatial memories triggers repulsion of hippocampal representations. *Current Biology*, *27*(15), 2307-2317.

Chanales, A. J., Tremblay-McGaw, A. G., Drascher, M. L., & Kuhl, B. A. (2021). Adaptive repulsion of long-term memory representations is triggered by event similarity. *Psychological science*, *32*(5), 705-720.

Chang, L., & Tsao, D. Y. (2017). The code for facial identity in the primate brain. *Cell*, 169(6), 1013–1028.

Chen, J., Leber, A. B., & Golomb, J. D. (2019). Attentional capture alters feature perception. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(11), 1443.

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature neuroscience*, 20(1), 115.

Chen, J. M., Norman, J. B., & Nam, Y. (2021). Broadening the stimulus set: introducing the American multiracial faces database. *Behavior Research Methods*, 53(1), 371-389.

Chen, P. H. C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., & Ramadge, P. J. (2015). A reduced-dimension fMRI shared response model. In *Advances in Neural Information Processing Systems* (pp. 460-468).

Chung, K. M., Kim, S., Jung, W. H., & Kim, Y. (2019). Development and validation of the Yonsei face database (YFace DB). *Frontiers in psychology*, 10, 2626.

Chunharas, C., Brady, T., & Ramachandran, V. S. (2018). Selective amplification of salient features of visual memories during early memory consolidation. *PsyArXiv.* https://doi.org/10.31234/osf.io/5dcxa

Chunharas, C., Rademaker, R. L., Brady, T., & Serences, J. (2019). Adaptive memory distortion in visual working memory. *PsyArXiv.*

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature neuroscience*, 17(3), 455-462.

Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., Norman, K.A., Pillow, J., Ramadge, P.J., Turk-Browne, N.B., & Willke, T. L. (2017). Computational approaches to fMRI analysis. *Nature neuroscience*, 20(3), 304-313.

Colgin, L. L., Moser, E. I., & Moser, M. B. (2008). Understanding memory through hippocampal remapping. *Trends in neurosciences*, 31(9), 469-477.

Conley, M. I., Dellarco, D. V., Rubien-Thomas, E., Cohen, A. O., Cervera, A., Tottenham, N., & Casey, B. J. (2018). The racially diverse affective expression (RADIATE) face stimulus set. *Psychiatry research*, 270, 1059-1067.

Cooper, R. A., Kensinger, E. A., & Ritchey, M. (2019). Memories fade: The relationship between memory vividness and remembered visual salience. *Psychological science*, *30*(5), 657-668.

Cooper, R. A., & Ritchey, M. (2019). Cortico-hippocampal network connections support the multidimensional quality of episodic memory. *Elife*, *8*, e45591.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 23(6), 681–685.

Cowen, A. S., Chun, M. M., & Kuhl, B. A. (2014). Neural portraits of perception: reconstructing face images from evoked brain activity. *Neuroimage*, 94, 12-22.

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 10(4-5), 171-178.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI)“brain reading”: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*, 19(2), 261-270.

Crowder, R. G. (2014). The interference theory of forgetting in long-term memory. In *Principles of Learning and Memory* (pp. 234-279). Psychology Press.

Dado, T., Güçlütürk, Y., Ambrogioni, L., Ras, G., Bosch, S., van Gerven, M., & Güçlü, U. (2022). Hyperrealistic neural decoding for reconstructing faces from fMRI activations via the GAN latent space. *Scientific reports*, 12(1), 1-9.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*, 9(2), 179-194.

Davis, T., & Poldrack, R. A. (2013). Measuring neural representations with fMRI: practices and pitfalls. *Annals of the New York Academy of Sciences*, 1296(1), 108-134.

DeBruine, Lisa; Jones, Benedict (2017): Face Research Lab London Set. figshare. Dataset. https://doi.org/10.6084/m9.figshare.5047666.v5

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of experimental psychology*, 58(1), 17.

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage*, 53(1), 1-15.

Diana, R. A., Peterson, M. J., & Reder, L. M. (2004). The role of spurious feature familiarity in recognition memory. *Psychonomic bulletin & review*, 11(1), 150-156.

Drascher, M. L., & Kuhl, B. A. (2022). Long-term memory interference is resolved via repulsion and precision along diagnostic memory dimensions. *Psychonomic Bulletin & Review*, 1-15.

Drucker, D. M., & Aguirre, G. K. (2009). Different spatial scales of shape similarity representation in lateral and ventral LOC. Cerebral Cortex, 19(10), 2269-2280.

Ebner, N. C., Riediger, M., and Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42(1):351–362.

Edwards, G. J., Cootes, T. F., & Taylor, C. J. (1998). Face recognition using active appearance models. In *European conference on computer vision*. Springer, Berlin, Heidelberg, pp. 581–595.

Engell, A. D., & Haxby, J. V. (2007). Facial expression and gaze-direction in human superior temporal sulcus. *Neuropsychologia*, 45(14), 3234-3241.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature methods*, 16(1), 111-116.

Ester, E. F., Sprague, T. C., & Serences, J. T. (2015). Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. *Neuron*, 87(4), 893-905.

Favila, S. E., Chanales, A. J. H., & Kuhl, B. A. (2016). Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nature Communications*, 7(1), 11066.

Favila, S. E., Kuhl, B. A., & Winawer, J. (2022). Perception and memory have distinct spatial tuning properties in human visual cortex. *Nature communications*, 13(1), 1-21.

Favila, S. E., Samide, R., Sweigart, S. C., & Kuhl, B. A. (2018). Parietal representations of stimulus features are amplified during memory retrieval and flexibly aligned with top-down goals. Journal of *Neuroscience*, 38(36), 7809-7821.

Fawcett, J. M., & Hulbert, J. C. (2020). The many faces of forgetting: Toward a constructive view of forgetting in everyday life. *Journal of Applied Research in Memory and Cognition*, *9*(1), 1-18.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), 1.

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., & Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, (47), S102.

Ford, J. H., & Kensinger, E. A. (2016). Effects of internal and external vividness on hippocampal connectivity during memory retrieval. *Neurobiology of learning and memory*, 134, 78-90.

Furl, N., Henson, R. N., Friston, K. J., & Calder, A. J. (2013). Top-down control of visual responses to fear by the amygdala. *Journal of Neuroscience*, 33(44), 17435-17443.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological review*, 91(1), 1.

Goldstone, R. L. (1998). Perceptual learning. *Annual review of psychology*, *49*(1), 585-612.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, 78(1), 27-43.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of experimental psychology: General*, *130*(1), 116.

Golomb, J. D. (2015). Divided spatial attention and feature-mixing errors. *Attention, Perception, & Psychophysics*, *77*(8), 2562-2569.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, 13.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, 48(1), 63-72.

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., & van Gerven, M. A. (2017). Reconstructing perceived faces from brain activations with deep adversarial neural decoding. *Advances in neural information processing systems*, 30.

Harlow, I. M., & Donaldson, D. I. (2013). Source accuracy data reveal the thresholded nature of human episodic memory. *Psychonomic Bulletin & Review*, *20*(2), 318-325.

Harlow, I. M., & Yonelinas, A. P. (2016). Distinguishing between the success and precision of recollection. *Memory*, *24*(1), 114-127.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425-2430.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., Hanke, M., & Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2), 404-416.

Hays, J., Wong, C., & Soto, F. A. (2020). FaReT: A free and open-source toolkit of three-dimensional models and software to study face perception. *Behavior research methods*, 52(6), 2604-2622.

Haynes, J. D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature neuroscience*, 8(5), 686-691.

Horner, A. J., & Burgess, N. (2013). The associative structure of memory for multi-element events. *Journal of Experimental Psychology: General*, 142(4), 1370.

Horner, A. J., & Burgess, N. (2014). Pattern completion in multielement event engrams. *Current Biology*, 24(9), 988-992.

Hulbert, J. C., & Norman, K. A. (2015). Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice. *Cerebral Cortex*, *25*(10), 3994-4008.

Hutchinson, J. B., Pak, S. S., & Turk-Browne, N. B. (2016). Biased competition during long-term memory formation. *Journal of cognitive neuroscience*, 28(1), 187-197.

Huth, A. G., Lee, T., Nishimoto, S., Bilenko, N. Y., Vu, A. T., & Gallant, J. L. (2016). Decoding the semantic content of natural movies from human brain activity. *Frontiers in systems neuroscience*, 10, 81.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2), 825-841.

Jiang, Z., Sanders, D. M. W., & Cowell, R. A. (2022). Visual and semantic similarity norms for a photographic image stimulus set containing recognizable objects, animals and scenes. *Behavior Research Methods*, 1-17.

Kahana, M. J., Zhou, F., Geller, A. S., & Sekuler, R. (2007). Lure similarity affects visual episodic recognition: Detailed tests of a noisy exemplar model. *Memory & cognition*, 35(6), 1222-1232.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5), 679.

Kamitani, Y., & Tong, F. (2006). Decoding seen and attended motion directions from activity in the human visual cortex. *Current biology*, 16(11), 1096-1102.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature neuroscience*, 3(8), 759-763.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(7185), 352-355.

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Chaibub Neto, E., & Keshavan, A. (2017). Mindboggling morphometry of human brains*. PLoS computational biology*, 13(2), e1005350.

Kleiner, M. Brainard, D., & Pelli, D. (2007) What's new in Psychtoolbox-3? *Perception*, 36 (ECVP Abstract Supplement), 14.

Koestinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2011, November). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)* (pp. 2144-2151). IEEE.

Korkki, S. M., Richter, F. R., Jeyarathnarajah, P., & Simons, J. S. (2020). Healthy ageing reduces the precision of episodic memory retrieval. *Psychology and Aging*, 35(1), 124.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, 93(3), 480-490.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 4.

Kroon, D.J. (2012). Active Shape Model (ASM) and Active Appearance Model (AAM). *MATLAB Central File Exchange*.

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*(2), 225-248.

Kuhl, B. A., & Chun, M. M. (2014). Successful remembering elicits event-specific activity patterns in lateral parietal cortex. *Journal of Neuroscience*, 34(23), 8051-8060.

Kuhl, B. A., Rissman, J., Chun, M. M., & Wagner, A. D. (2011). Fidelity of neural reactivation reveals competition between memories. *Proceedings of the National Academy of Sciences*, 108(14), 5903–5908.

LaBar, K. S., Crupain, M. J., Voyvodic, J. T., & McCarthy, G. (2003). Dynamic perception of facial affect and identity in the human brain. *Cerebral Cortex*, 13(10), 1023-1033.

Lakshmi, A., Wittenbrink, B., Correll, J., & Ma, D. S. (2021). The India face set: International and cultural boundaries impact face impressions and perceptions of category membership. *Frontiers in psychology*, 12, 627678.

Lee, H., & Kuhl, B. A. (2016). Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex. *Journal of Neuroscience*, 36(22), 6069-6082.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature neuroscience*, 4(1), 89-94.

Li, A. Y., Fukuda, K., Lee, A. C., & Barense, M. D. (2020). Visual interference can help and hinder memory: Capturing representational detail using the Validated Circular Shape Space. *bioRxiv*, 535922.

Lin, P. H., & Luck, S. J. (2009). The influence of similarity on visual working memory representations. *Visual Cognition*, 17(3), 356-372.

Long, N. M., & Kuhl, B. A. (2018). Bottom-up and top-down factors differentially influence stimulus representations across large-scale attentional networks. *Journal of Neuroscience*, 38(10), 2495-2504.

Ma, D. S., Correll, J., and Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47(4):1122–1135.

Ma, D. S., Kantner, J., & Wittenbrink, B. (2021). Chicago face database: Multiracial expansion. *Behavior Research Methods*, 53(3), 1289-1300.

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, 113(46), 13203-13208.

Mate, J., & Baqués, J. (2009). Short article: Visual similarity at encoding and retrieval in an item recognition task. *Quarterly Journal of Experimental Psychology*, 62(7), 1277-1284.

Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *elife*, 7, e31873.

McClelland, J. L., McNaughton, B. L., & O'reilly, R. C. (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. *Psychological review*, 102(3), 419.

Melton, A. W., & Irwin, J. M. (1940). The influence of degree of interpolated learning on retroactive inhibition and the overt transfer of specific responses. *The American Journal of Psychology*, 53(2), 173-203.

Milborrow, S., Morkel, J., & Nicolls, F. (2010). The MUCT landmarked face database. *Pattern recognition association of South Africa*, 201(0).

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior research methods, instruments, & computers*, 36(4), 630-633.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M. A., Morito, Y., Tanabe, H. C., Sadato, N., & Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5), 915-929.

Mozafari, M., Reddy, L., & VanRullen, R. (2020, July). Reconstructing natural scenes from fMRI patterns using BigBiGAN. In *2020 international joint conference on neural networks (IJCNN)* (pp. 1-8). IEEE.

Muller, R. U., & Kubie, J. L. (1987). The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *Journal of Neuroscience*, 7(7), 1951-1968.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of classification*, 31(3), 274-295.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6), 902-915.

Nemrodov, D., Behrmann, M., Niemeier, M., Drobotenko, N., & Nestor, A. (2019). Multimodal evidence on shape and surface information in individual face processing. *Neuroimage*, 184, 813-825.

Nestor, A., Lee, A. C., Plaut, D. C., & Behrmann, M. (2020). The face of image reconstruction: progress, pitfalls, prospects. *Trends in cognitive sciences*, 24(9), 747-759.

Nestor, A., Plaut, D. C., & Behrmann, M. (2016). Feature-based face representations and image reconstruction from behavioral and neural data. *Proceedings of the National Academy of Sciences*, 113(2), 416-421.

Nilakantan, A. S., Bridge, D. J., Gagnon, E. P., VanHaerents, S. A., & Voss, J. L. (2017). Stimulation of the posterior cortical-hippocampal network enhances precision of memory recollection. *Current Biology*, *27*(3), 465-470.

Nilakantan, A. S., Bridge, D. J., VanHaerents, S., & Voss, J. L. (2018). Distinguishing the precision of spatial recollection from its success: Evidence from healthy aging and unilateral mesial temporal lobe resection. *Neuropsychologia*, *119*, 101-106.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19), 1641-1646.

Norman, K. A. (2010). How hippocampus and cortex contribute to recognition memory: revisiting the complementary learning systems model. *Hippocampus*, 20(11), 1217-1227.

Norman, K. A., Newman, E. L., & Detre, G. (2007). A neural network model of retrieval-induced forgetting. *Psychological review*, *114*(4), 887.

Norman, K. A., Newman, E., Detre, G., & Polyn, S. (2006). How inhibitory oscillations can train neural networks and punish competitors. *Neural computation*, *18*(7), 1577-1610.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9), 424-430.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological review*, 110(4), 611.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087-11092.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, 4(6), 661-682.

O'Reilly, R. C., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in cognitive sciences*, 6(12), 505-510.

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychological review*, 108(2), 311.

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

Pertzov, Y., Manohar, S., & Husain, M. (2017). Rapid forgetting results from competition over time between items in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(4), 528.

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, 119(17), e2115228119.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage*, 84, 320-341.

Rajsic, J., Swan, G., Wilson, D. E., & Pratt, J. (2017). Accessibility limits recall from visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(9), 1415.

Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: a robust approach. *Neuroimage*, 53(4), 1181-1196.

Rhodes, S., Abbene, E. E., Meierhofer, A. M., & Naveh-Benjamin, M. (2020). Age differences in the precision of memory at short and long delays. *Psychology and Aging*, *35*(8), 1073.

Richter, F. R., Cooper, R. A., Bays, P. M., & Simons, J. S. (2016). Distinct neural mechanisms underlie the success, precision, and vividness of episodic memory. *Elife*, *5*, e18260.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, 21(4), 803.

Roesch, E. B., Tamarit, L., Reveret, L., Grandjean, D., Sander, D., & Scherer, K. R. (2011). FACSGen: A tool to synthesize emotional facial expressions through systematic manipulation of facial action units. *Journal of Nonverbal Behavior*, 35(1), 1-16.

Rugg, M. D., Otten, L. J., & Henson, R. N. (2002). The neural basis of episodic memory: evidence from functional neuroimaging. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1424), 1097-1110.

Rundus, D. (1973). Negative effects of using list items as recall cues. *Journal of Verbal Learning and Verbal Behavior*, 12(1), 43-50.

Said, C. P., Moore, C. D., Engell, A. D., Todorov, A., & Haxby, J. V. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of vision*, 10(5), 11-11.

Sanocki, T., & Sulman, N. (2011). Color relations increase the capacity of visual short-term memory. *Perception*, 40(6), 635-648.

Sato, W., Yoshikawa, S., Kochiyama, T., & Matsumura, M. (2004). The amygdala processes the emotional significance of facial expressions: an fMRI investigation using the interaction between expression and face direction. *Neuroimage*, 22(2), 1006-1013.

Schacter, D. L., Guerin, S. A., & Jacques, P. L. S. (2011). Memory distortion: an adaptive perspective. *Trends in cognitive sciences*, 15(10), 467-474.

Schacter, D. L., & Madore, K. P. (2016). Remembering the past and imagining the future: Identifying and enhancing the contribution of episodic memory. *Memory Studies*, 9(3), 245-255.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711), 20160049.

Schindler, S., Zell, E., Botsch, M., & Kissler, J. (2017). Differential effects of face-realism and emotion on event-related brain potentials and their implications for the uncanny valley theory. *Scientific reports*, 7(1), 1-13.

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature communications*, 6, 8151.

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature human behaviour*, 4(11), 1156-1172.

Scotti, P. S., Hong, Y., Golomb, J. D., & Leber, A. B. (2021). Statistical learning as a reference point for memory distortions: Swap and shift errors. *Attention, Perception, & Psychophysics*, 1-21.

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., & van Gerven, M. A. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181, 775-785.

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological science*, 20(2), 207-214.

Shen, B., RichardWebster, B., O'Toole, A., Bowyer, K., & Scheirer, W. J. (2021, December). A study of the human perception of synthetic faces. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition* (FG 2021) (pp. 1-8). IEEE.

Smith, R. E., & Hunt, R. R. (2000). The influence of distinctive processing on retrieval-induced forgetting. *Memory & Cognition*, *28*(4), 503-508.

Steyvers, M. (1999). Morphing techniques for manipulating face images. *Behavior Research Methods, Instruments, & Computers*, 31(2), 359-369.

St-Laurent, M., Abdi, H., & Buchsbaum, B. R. (2015). Distributed patterns of reactivation predict vividness of recollection. *Journal of Cognitive Neuroscience*, 27(10), 2000-2018.

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2008). Accelerated relearning after retrieval-induced forgetting: the benefit of being forgotten. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1), 230.

Sun, S. Z., Fidalgo, C., Barense, M. D., Lee, A. C., Cant, J. S., & Ferber, S. (2017). Erasing and blurring memories: The differential impact of interference on separate aspects of forgetting. *Journal of Experimental Psychology: General*, *146*(11), 1606.

Swan, G., Collins, J., & Wyble, B. (2016). Memory for a single object has differently variable precisions for relevant and irrelevant features. *Journal of vision*, 16(3), 32-32.

Theves, S., Fernández, G., & Doeller, C. F. (2020). The hippocampus maps concept space, not feature space. *Journal of Neuroscience*, *40*(38), 7318-7325.

Thomas, K. M., Drevets, W. C., Whalen, P. J., Eccard, C. H., Dahl, R. E., Ryan, N. D., & Casey, B. J. (2001). Amygdala response to facial expressions in children and adults. *Biological psychiatry*, 49(4), 309-316.

Tompary, A., & Thompson-Schill, S. L. (2021). Semantic influences on episodic memory distortions. *Journal of Experimental Psychology: General*.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D.J., Westerlund, A., Casey, B.J., & Nelson, C. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry research*, 168(3), 242-249.

Tulving, E. (1974). Cue-dependent forgetting: When we forget something we once knew, it does not necessarily mean that the memory trace has been lost; it may only be inaccessible. *American scientist*, 62(1), 74-82.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1), 71-86.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging*, 29(6), 1310-1320.

Van Ginneken, B., Frangi, A. F., Staal, J. J., ter Haar Romeny, B. M., & Viergever, M. A. (2002). Active shape model segmentation with optimal features. *IEEE transactions on medical imaging*, 21(8), 924-933.

VanRullen, R., & Reddy, L. (2019). Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications biology*, 2(1), 1-10.

Walker, M., Schönborn, S., Greifeneder, R., & Vetter, T. (2018). The Basel Face Database: A validated set of photographs reflecting systematic differences in Big Two and Big Five personality dimensions. *PloS one*, 13(3), e0193190.

Wanjia, G., Favila, S. E., Kim, G., Molitor, R. J., & Kuhl, B. A. (2021). Abrupt hippocampal remapping signals resolution of memory interference. *Nature communications*, 12(1), 1-11.

Watson, H. C., & Lee, A. C. (2013). The perirhinal cortex and recognition memory interference. *Journal of Neuroscience*, 33(9), 4192-4200.

Wen, H., Shi, J., Zhang, Y., Lu, K. H., Cao, J., & Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 28(12), 4136-4160.

Wheatley, T., Weinberg, A., Looser, C., Moran, T., & Hajcak, G. (2011). Mind perception: Real but not artificial faces sustain neural activity beyond the N170/VPP. *PloS one*, 6(3), e17960.

Wills, T. J., Lever, C., Cacucci, F., Burgess, N., & O'keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. *Science*, 308(5723), 873-876.

Won, B. Y., Haberman, J., Bliss-Moreau, E., & Geng, J. J. (2020). Flexible target templates improve visual search accuracy for faces depicting emotion. *Attention, Perception, & Psychophysics*, 1-15.

Xue, G. (2018). The neural representations underlying human episodic memory. *Trends in Cognitive Sciences*, 22(6), 544-561.

Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater neural pattern similarity across repetitions is associated with better memory. *Science*, 330(6000), 97-101.

Yassa, M. A., & Stark, C. E. (2011). Pattern separation in the hippocampus. *Trends in neurosciences*, 34(10), 515-525.

Yeung, L. K., Ryan, J. D., Cowell, R. A., & Barense, M. D. (2013). Recognition memory impairments caused by false recognition of novel objects. *Journal of Experimental Psychology: General*, 142(4), 1384.

Yu, X., & Geng, J. J. (2019). The attentional template is shifted and asymmetrically sharpened by distractor context. *Journal of experimental psychology: human perception and performance*, *45*(3), 336.

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*, 75(1), 168-179.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1), 45-57.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, 453(7192), 233-235.

Zhao, Y., Chanales, A. J., & Kuhl, B. A. (2021). Adaptive memory distortions are predicted by feature representations in parietal cortex. *Journal of Neuroscience*, *41*(13), 3014-3024.

Zheng, J., Schjetnan, A. G., Yebra, M., Gomes, B. A., Mosher, C. P., Kalia, S. K., Valiante, T.A., Mamelak, A.N., Kreiman, G., & Rutishauser, U. (2022). Neurons detect cognitive boundaries to structure episodic memories in humans. *Nature Neuroscience*, 25(3), 358-368.