SENSORY HISTORY MATTERS FOR VISUAL REPRESENTATION:

IMPLICATIONS FOR AUTISM

David Alexander Kahn

A DISSERTATION

in

Neuroscience

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation	Co-Supervisor of Dissertation
Robert T. Schultz	Geoffrey Karl Aguirre
Professor of Psychology	Associate Professor of Neurology

Graduate Group Chairperson

Joshua Gold, Professor of Neuroscience

Dissertation Committee:

Russell Epstein (Chair)	Professor of Psychology
David Brainard	RRL Professor of Psychology
James Loughead	Research Associate Professor of Psychology in Psychiatry
Nicole Rust	Assistant Professor of Psychology

SENSORY HISTORY MATTERS FOR VISUAL REPRESENTATION: IMPLICATIONS FOR AUTISM

Acknowledgements

I was fortunate to be guided in my graduate studies by two advisors, Bob Schultz and Geoff Aguirre. With each, I built an intellectual foundation from which a project bridging their disparate interests could be launched. Together, they fostered an environment full of theory, pragmatism, and patience for me to grow as a scientist. Somehow they each knew exactly the amount of guidance I needed, and also when it was best to let me wander. I'm beyond thankful for their mentorship.

I would like to thank the members of my thesis committee, Russell Epstein, David Brainard, Nicole Rust, & James Loughead, for their insight, patience, & assistance.

I am grateful to many mentors and colleagues from my time at Caltech for guiding my development as a young academic. Ralph Adolphs was largely responsible for kindling my interest in cognitive neuroscience and opened his door to an inexperienced student. I aspire to his steady and joyful approach to science and life. My first advisor, Marianne Bronner, is enduringly a role model of professorship and lab leadership. When a lab community just feels right to me now, it is because it reminds me of hers. In addition, Lisa Taneyhill, Mike Tyszka, Michael Spezio, Lynn Paul, and Kevin Gilmartin were warm guides and great friends.

I would like to thank the many members of the Aguirre Lab at Penn, the Center for Autism Research at the Children's Hospital of Philadelphia, and the broader Penn community for their friendship and assistance. In particular, Lisa Guy & Ashley de Marchena, who performed the clinical assessments the the final study in this thesis, were unparalleled guides in the nuances of clinical psychology. Lori Le made up for my shortcomings with scheduling and administration; without her the final study in this thesis would have stood still. Dave Wolk generously instructed me early on in electroencephalography and offered the use of his lab for data collection. Ruben Gur was my first rotation advisor at Penn and encouraged me to trust and follow my interests; I always will. Marcelo Mattar collaborated on the unified framework of stimulus history effects. Many of the ideas, including the mapping of effects across the cortical surface, came about in really fun conversations with him. Alison Harris and I worked closely on the first study in this thesis. Most of my knowledge of the evoked responses to faces and my ability with crossword puzzles I owe to her.

Thanks to everyone at the Penn Neuroscience Graduate Group. For their advice & guidance I'd like to thank the three NGG chairs during my time at Penn: Mikey Nusbaum, Rita Balice-Gordon, and Josh Gold. Jane Hoshi was the most able coordinator; and endless source of help and calm. My fellow students were the most supportive, inspiring, and fun group; I'm thrilled to have found them as friends and colleagues.

This work would not have been possible without the contribution of the many participants and families of participants who provided data for these studies. I am thankful for their patience, enthusiasm, and definitely their attention.

My work was supported by many sources of funding, including grants from the National Institutes of Health (NIH R21-MH98153-01, NIH K08 MH72926-01, NIH T32-

MH17168-30), the National Science Foundation (SBE-0542013) and the Burroughs-Wellcome Fund.

I have been blessed with so many close friends. Thanks to Michael & Molly, Doug, Damien, Laura, Jessie, Kate, Andy & Catherine, Mark & Cori, Dorota, Mike, Peter, Bret, Mary, Sean, Brandon, Emilia, Jesse, Mike & Teddy, Will, Kate, Laura, Vanessa, Lindsay, Toni, Sam, Alana, and Rob & the Hawks.

My siblings are the steadiest sources of inspiration, humor, and love in my life. Thank you Sam, Winston, Margot, & Scott.

This dissertation is dedicated to my parents, Karyn and Shepard. You guys are the best.

ABSTRACT

SENSORY HISTORY MATTERS FOR VISUAL REPRESENTATION: IMPLICATIONS FOR AUTISM

David Alexander Kahn

Robert T. Schultz

Geoffrey Karl Aguirre

How does the brain represent the enormous variety of the visual world? An approach to this question recognizes the types of information that visual representations maintain. The work in this thesis begins by investigating the neural correlates of perceptual similarity & distinctiveness, using EEG measurements of the evoked response to faces. In considering our results, we recognized that the effects being measured shared intrinsic relationships, both in measurement and in their theoretic basis. Using carry-over fMRI designs, we explored this relationship, ultimately demonstrating a new perspective on stimulus relationships based around sensory history that best explains the modulation of brain responses being measured. The result of this collection of experiments is a unified model of neural response modulation based around the integration of recent sensory history into a continually-updated reference; a "drifting-norm."

With this novel framework for understanding neural dynamics, we tested whether cognitive theories of autism spectrum disorder (ASD) might have a foundation in altered neural coding for perceptual information. Our results suggest ASD brain responses

depend on a more moment-to-moment understanding of the visual world relative to neurotypical controls. This application both provides an exciting foothold in the brain for future investigations into the etiology of ASD, and validates the importance of sensory history as a dimension of visual representation.

Contents

List of Tables

List of Figures

1 Introduction

Somewhere toward the middle of my PhD work, my younger brother Sam and I were discussing the visual system over the phone. Having studied film as an undergrad, he brought up a simple example of how the visual system doesn't work like a camera. He noted that when one tries to film a sunset with a manual film camera and a fixed aperture, the recorded scene on film would rapidly fall to black, whereas the human eye would experience a longer, gradual fade - continuing after the sun had fallen from view. The reasons for this are many, but the biggest one is that the eye doesn't maintain a fixed aperture; it automatically adjusts to changes in light intensity to maintain sensitivity. This dynamic adjustment contains one implicit decision of the visual system - opting for sensitivity over an accurate representation of physical reality.

On a breezy day in early February 2012, my close friend Dorota & I visited an installation by the artist Doug Wheeler entitled SA MI 75 DZ NY 12 at David Zwirner in New York. The installation is what Wheeler calls an "infinity environment." It's difficult to report exactly how large the environment is - perhaps 50 feet in diameter and circular in shape. Inside, the floor and ceiling curve to meet the walls in (what one presumes) are parabolic curves. The paint is consistent & diffusely reflective, and the whole space is lit from the edges of the vestibule through which one walks to enter. With no edges or textures, with one's back to the entrance there is simply nothing for the eye to fixate on inside. The effect is something akin to being in a fog, but with a greater (yet ambiguous) sense of depth. There is simply nothing to see.

…Except that there is. After a few minutes inside the space, Dorota walked over to me & expressed amazement at how many “floaters” were in her eyes. I too, had never seen so many of them. The percept is a side effect of anatomy; the vitreous humour of the eye breaks down over time & the resulting imperfections can cast shadows on the retina. These shadows are always there. When there’s not much else to see, for instance when watching the sky on a clear day, it’s hard not to perceive them. But even as I write this and look around the library where I am siting, I struggle to perceive any. I would suggest this is another implicit decision of the visual system: when there is plenty to “see”, the system ignores some things.

These are just two implicit decisions of the visual system. In the process of perception, countless such trade-offs occur.

Coming back to the camera’s perspective, the history of computer vision (up until last year) has demonstrated another issue: making sense of image data is really hard. Consider the hurdles in creating an algorithm to recognize your car or bicycle using a camera. It must take into account every color spectra the paint could reflect under different illuminations and every angle of view. Perhaps with enough work, the algorithm one day seems pretty good… until you park behind a chain link fence. Or (to make the analogy even more explicit) a smudge gets on the camera lens. From this example we realize that *ignoring* some information can facilitate other functions.

These observations are central to theories of perceptual systems that focus on the nature of neural information (Barlow, 1961; Olshausen & Field, 2004). The crux of many such theories is that sensory information is processed into *neural codes* that manage to exploit features of the information in adaptive ways. The corner of vision science this thesis finds itself in is focused on these neural codes. What information is stored in them, and what is the relationship between different types of information? How do we measure these codes in the brain? And could we possibly link differences in these brain measurements to changes in behavior, or to psychiatric disorders? The arc of this thesis reframes two types of information stored in visual neural codes in terms of one unified model. The final chapter applies this unified perspective to look for neural coding differences in autism spectrum disorder.

How does one begin to approach the question of information encoded in the brain? One of the most useful tools is derived from psychology - the concept of a *perceptual space* or *stimulus space*. A perceptual space is a theoretic multidimensional space in which all manner of stimuli (in our case, visual objects such as faces and cars etc.) can be placed. The dimensions of this space allow for mapping and measuring - and thus for understanding the visual relationships between all elements in the space. For simplicity, we will reduce the size of the space under discussion to just the space of a special class of objects: faces. A perceptual space for faces would need to be quite highly dimensioned; though faces are amazingly similar, there is still a wonder of variety between and within individuals. Consider what would be necessary to represent just my face and the faces of my two brothers, Winston and Sam. I’m often told we look quite similar; if one assumes that each brother is equally similar perceptually to each other, the space of our faces could be represented as a plane with three points forming an equilateral triangle. Of course we're never quite making the same expression; perhaps my smiling face is more like my younger brother's smiling face than his frowning face. We will need to add a dimension to account for this. And of course sometimes we let our beards grow out a bit: more variation, another dimension. One can see how quickly perceptual spaces can grow! The useful thing about the perceptual space is that they reduce just as easily - indeed we started by reducing to just the space of faces, and then just the space of 3 brothers. Most of the work in this thesis starts with very reduced perceptual spaces - a single vector in "face-space" between two face identities. Having discussed the perceptual space as a tool, the appeal of this simplicity should be intuitive to the reader. (It is important to keep in mind while reading that any understanding we derive here will need to be studied eventually in respect to how it scales).

What does the perceptual space as a tool offer the researcher? Like any map, at its simplest, it offers distances. These distances in perceptual space can be thought of as indexes of *perceptual similarity.* Psychologists have put these maps and distances to great use for years. To use them to investigate neural codes however, we also require a yardstick for the brain - some measure of distance in neural space. A few such yardsticks exist. The one most heavily exploited in this work is called *neural adaptation.* Adaptation is a phenomenon by which the repeated presentation of a stimulus will yield an attenuated neural response. This has been observed using neuroimaging measures such as functional magnetic resonance imaging (fMRI, Grill-Spector & Malach, 2001) and single-cell recordings (Leopold *et al*., 2006), among other methods. Work by one of my advisors, Geoffrey Aguirre, just prior to the initiation of this thesis sought to extend this work beyond identical repetitions to investigate where and how this neural yardstick was neatly notched for distance (Aguirre, 2007; Drucker & Aguirre, 2009). The early work (Kahn *et al.,* 2010) presented in this thesis uses neural adaptation as a yardstick for perceptual distance within the evoked response to faces measured by electroencephalography (EEG).

(During the period during which this these was performed, other yardsticks came into use. One of the most exciting uses distributed patterns of activity in fMRI to index similarity. By evaluating the correlations between distributed activity evoked by different stimuli across a region of cortex in fMRI, it is possible to index the similarity between neural representations. This representational similarity analysis (RSA, Kriegeskorte *et al.*, 2009) and other types of multi-voxel pattern analysis (MVPA) share a complex relationship with neural adaptation (Epstein & Morgan, 2012)).

Returning to our perceptual space: distance, some have argued, is not the only tool the space offers to us. It has been proposed that a perceptual space, such as that of faces, must have a center - or *norm* - that might have unique qualities. The idea of a *norm-based* code is that coordinates in neural space are referenced to the norm rather than merely to each other. By way of analogy, one could locate my hometown of Cleveland, OH by the coordinates 41˚28’56”N 81˚40’11”W. These coordinate place particular emphasis on a norm - the intersection of the equator and the prime meridian. (A separate method of locating Cleveland would be to offer a collection of distances from other cities; 190 miles from Toronto, 408 miles from New York City, 311 miles from Chicago). Psychological & psychometric investigations of perceptual norms suggested some special properties likely exist. One class of such findings is adaptive *after-effects* (Leopold *et al.*, 2001), in which exposure to a particular exemplar to one extreme of a norm could bias the norm in the opposite ("anti") direction. Others properties were derived from the observation that norms were used a reference more often than they were referred to other exemplars. A classic example (Tversky, 1977) is that 99 is judged as more similar to 100 than 100 is to 99. Yet another is the finding that average stimuli are judged as more familiar than extreme exemplars. For instance, when a collection of extreme exemplars from a simple cartoon face space is presented repeatedly, a familiarity bias is induced for the norm even if it was never seen; viewers will judge the norm as more likely seen relative to an extreme exemplar that was actually presented (Posner & Keele, 1968).

In a perceptual space, the norm might not offer a ready yardstick but rather something more like a compass. Just prior to and during the early years of this thesis project, researchers using fMRI sought to find neural evidence of just such a compass. The hypothesized neural correlate was a reduced level of bulk neural response for a norm relative to an extreme exemplar. Studies investigating faces (Loffler *et al*. 2005), face silhouettes (Davidenko *et al.*, 2011), and abstract shapes (Panis *et al.*, 2010) indexed these responses. In addition to the possibility of a differential amplitude of neural response, we hypothesized that norms might induce neural biases, similar to psychological biases of similarity. Just as an ellipse is judged as more like a circle than a circle is like an ellipse, we suspected our neural yardstick might measure differently when comparing more average stimuli to more extreme ones.

The first chapter of this thesis began to play with these tools - neural adaptation & norm-based effects - in a single EEG experiment. Though this experiment, we began to reconsider some of the seemingly straightforward definitions offered above. What we discovered was the interactions between effects were hard to disentangle. The following two chapters continue this reassessment. What coalesces is a new perspective on neural codes, yardsticks, and visual perception.

The final chapter of this thesis presents recent work applying this new perspective to an investigation of the neural etiology of autism spectrum disorder (ASD). ASD is a heterogeneous group of developmental disorders clinical defined by social communication deficits and a tendency toward restricted interests and repetitive behaviors (DSM-V - APA, 2013).

One of the oft-cited characteristics of ASD is an acute processing of detail in everyday experience. Theoretical approaches of ASD have sought to explain this tendency from a neurocognitive standpoint. One the most enduring theories is the weak central coherence account (Frith, 1989; Happé & Frith, 2006), which highlighted a difficulty to extract global meaning from features in ASD, likely driven by a bias for local-level information. Parallels are often drawn between weak central coherence and the stereotyped ASD cognitive style of “missing the forest for the trees.” Related theories highlighted other imbalances: the “enhanced discrimination and reduced generalization” hypothesis (Plaisted, 2001) focused on a differential manifestation of perceptual ability, as did the “enhanced perceptual functioning” account (Mottron *et al*. 2006), which echoed the local-bias noted by weak central coherence. (I should note these theories form only one corner of the literature on altered cognitive functioning in ASD. A separate wing focuses on social-first theories of the disorder (e.g. Chevallier *et al*. 2012)).

A collection of findings have lent support to these theories: the demonstration of enhanced perceptual discrimination (Plaisted *et al*. 1998; O’Riordan & Plaisted, 2001), a reduced ability to generalize prototypes (Klinger & Dawson, 2001), and alterations in the face adaptive after-effects described earlier (Pellicano *et al*., 2007).

Early in this introduction, the point was raised that one of the trade-offs of sensory systems is the implicit choice to ignore certain information when there is already “enough.” From just the briefest description of these theories of ASD, it does not seem a stretch to speculate that just such a sensory trade-off (for example, implicitly deciding when to “ignore”) is differently balanced in ASD. This places the locus of dysfunction in ASD within the perceptual system, and likely within the nature of the neural codes.

The idea of altered neural coding driving the autistic phenotype has been proposed before. Two neural network theories of ASD were influential in the development of the final project in this thesis. McClelland (2000) summarizes a framework for cognitive development in which neural networks optimize neural codes dynamically. Within this framework, the central trade-off struck by these codes is between preserving generality (overlap or sharing between codes) or emphasizing the conjunction of features. In McClelland’s example, the former would encode the visual qualities of ‘red’ and ‘square’ separately (and allow thus allow ‘red’ to be used for a fire hydrant as well) while the latter would encode ‘red square’ in conjunction. This latter style was dubbed “hyperspecificity” - an increasingly conjunctive, less broadly connected code for all visual information - and proposed as a neural scheme to explain ASD. A related theory was that of Gustafsson (1997) who proposed that cortical feature maps (the neural Legos that assemble perceptual spaces) could be “inadequate” in ASD. He proposed that columns (the building blocks of feature maps) might be more narrowly tuned for their preferred stimulus features, and perhaps that the map itself could be more fragmented (using many columns in place of one broader one).

When developing this final proposal for this thesis, these proposals seemed to raise a single question: if neural codes are altered in autism, shouldn’t we be able to measure those differences with our neural yardstick of adaptation? This was exactly the route we took when designing an investigation of neural codes and autism. However, as the following three chapters unfold, it will become clear that our understanding of our neural yardstick & compass shifted and merged. The new tool we end up with is both more complex and simpler, and proves useful for indexing perceptual differences in autism.

2 Temporally Distinct Neural Coding of Perceptual Similarity and Prototype Bias

2.1 Abstract

Psychological models suggest that perceptual similarity can be divided into geometric effects, such as metric distance in stimulus space, and non-geometric effects, such as stimulus-specific biases. We investigated the neural and temporal separability of these effects in a carry-over, event-related potential (ERP) study of facial similarity. By testing this dual effects model against a temporal framework of visual evoked components, we demonstrate that the behavioral distinction between geometric and non-geometric similarity effects is consistent with dissociable neural responses across the time course of face perception. We find an ERP component between the "face-selective" N170 and N250 responses (the "P200") that is modulated by transitions of face appearance, consistent with neural adaptation to the geometric similarity of face transitions. In contrast, the N170 and N250 reflect non-geometric stimulus bias, with different degrees of neural adaptation dependent upon the direction of transition within the stimulus space. These results suggest that the neural coding of perceptual similarity, in terms of both geometric and non-geometric representation, occurs rapidly and from relatively early in the perceptual processing stream.

2.2 Introduction

From searching for one's car in a parking lot to finding a friend in a crowd, we are confronted daily with varying exemplars from a given visual category. How does the visual system represent this variety? Several perceptual models are built around the notion of a “stimulus space,” a representation of comparative similarity based on observers’ judgments or their classification of stimuli into groups. Within-class stimulus variation may be mapped along the dimensions of this space. Rectangles, for instance, can be described in terms of aspect ratio and area, and color defined by variation in hue, saturation, and brightness.

A number of psychological models have related stimulus spaces to behavioral measures of perceptual similarity. So-called “geometric” models postulate a direct correspondence between the two, defining similarity in terms of the metric distance between two stimuli within a representational space (Shepard, 1964; Torgerson, 1965). While such geometric models are successful in explaining a wide range of behavior, certain perceptual properties of similarity violate these models (Holman, 1979; Krumhansel, 1978; Tversky, 1977). Notable is the violation of symmetry: while the ordering of a pair of stimuli should not alter their perceptual similarity in geometric models, this violation is frequently seen in practice. A classic perceptual example is that an ellipse is judged to be more similar to a circle than a circle is to an ellipse (Tversky, 1977). Often, such asymmetries suggest the existence of representational “prototypes” which can be interpreted as stimulus-specific biases producing non-geometric distortions of otherwise geometric similarity spaces. Prototypes may be the result of long-standing perceptual experience or the local effect of context induced by stimulus frequency (Polk, Behensky, Gonzalez, & Smith, 2002). Current models of similarity account for perceptual asymmetries through the inclusion of both geometric and non-geometric properties. The “additive similarity and bias” model of perceptual proximity (Holman, 1979; Nosofsky, 1991), for example, incorporates both geometric and non-geometric effects by defining the perceptual “proximity” of two stimuli as the sum of metric stimulus distance and stimulus bias, a term representing the stimulus-specific effects behind such asymmetries.

Supporting this distinction, studies of the neural representation of stimulus similarity have identified both geometric and non-geometric neural codes. A single-unit study of object perception demonstrated a correspondence between neural responsiveness in monkey inferotemporal cortex and the geometric organization of an abstract shape space, as derived from both behavioral and pixel-wise evaluations of similarity (Op de Beeck, Wagemans, & Vogels, 2001). Analogous geometric effects of similarity have been demonstrated in regions associated with object perception in humans using functional magnetic resonance imaging (fMRI; Drucker & Aguirre, 2009). Non-geometric similarity codes, in contrast, have been proposed to explain differential responsiveness to “prototypical” faces as compared to “distinctive” faces in fMRI (Loffler, Yourganov, Wilkinson, & Wilson, 2005).

Yet a great deal about the neural representation of perceptual similarity remains poorly understood. One major question relates to the dissociation of geometric and non-geometric effects at the neural level. While each of the studies cited above demonstrates neural correlates of either geometric or non-geometric encoding, no existing study has examined both types of effects concurrently. A second question is the time course of perceptual similarity effects: when in the perceptual processing stream do geometric and non-geometric coding of stimulus similarity occur? This latter question, extending to the temporal domain, speaks to the former by providing a non-spatial means of distinguishing these components of perceptual similarity.

In the present study, we investigated these questions using event-related potentials (ERPs). We hypothesized that geometric and non-geometric features of similarity would be evaluated during the time course of visual perception, and focused upon several of the early perceptual and “face-selective” components of the evoked visual response. In our analysis we examined four components of the ERP waveform previously associated with various stages of perceptual and mnemonic processing for faces. These include the P100, a marker of early visual processing (e.g., Di Russo, Martínez, Sereno, Pitzalis, & Hillyard, 2001), the N170 (occurring approximately 170 ms after stimulus onset) which is associated with perceptual encoding of the face (Bentin, Allison, Puce, & Perez, 1996; Itier & Taylor, 2004; Liu, Higuchi, Marantz, & Kanwisher, 2000; Sams, Hietanen, Hari, Ilmoniemi, & Lounasmaa, 1997), the P200, the positive component following the N170, and the N250, thought to reflect consolidation of perceptual representations into memory (Tanaka, Curran, Porterfield, & Collins, 2006). We used these components as elements of a temporal framework on which a neural model of geometric and non-geometric similarity effects could be evaluated.

We examined the sensitivity of this temporal framework to perceptual similarity by presenting faces varying in identity between two endpoint faces. Sensitivity to perceptual similarity was assessed via neural adaptation: a reduction in neural response following repeated stimulus presentation (Grill-Spector & Malach, 2001; Henson & Rugg, 2003). Previous work has demonstrated neural adaptation of “face-selective” responses in ERP (Jacques & Rossion, 2006; Itier & Taylor, 2002; Kovács et al., 2006) and the related methodology of magnetoencephalography, or MEG (Furl, van Rijsbergen, Treves, Friston, & Dolan, 2007; Harris & Nakayama, 2007; Harris & Nakayama, 2008). However, few of these studies have tested for parametric variation of adaptation effects, and the measurement of geometric and non-geometric similarity effects are often confounded. For example, while studies of prototype representation may observe differential response to centrally located stimuli (e.g., Loffler et al., 2005), these effects may result from the tendency of prototypical stimuli to be more similar to other stimuli and thus produce neural adaptation.

To disentangle these effects, we used a “carry-over design” (Aguirre, 2007) in which a continuous stream of stimuli is presented with first-order counterbalancing. The resulting data permit measurement of the direct effect of each stimulus upon the amplitude of neural response, as well as the modulatory effect of one stimulus upon the next (e.g., neural adaptation). Geometric neural similarity is revealed in this context as a symmetric, parametric adaptation of ERP response proportional to the change in perceptual similarity. Non-geometric neural similarity, suggestive of explicit neural representation of a prototype or central tendency of the stimulus space, was modeled as an asymmetric modulation of the ERP response dependent upon the direction of stimulus transition.

2.3 Materials and Methods

2.3.1 Subjects

Six right-handed subjects (3 women, 3 men) between the ages of 22 and 39 (mean age 29.5) with normal or corrected-to-normal vision participated in the study. All subjects provided informed consent under the guidelines of the Institutional Review Board of the University of Pennsylvania and the Declaration of Helsinki.

2.3.2 Stimuli

Two neutral faces (subtending 9.4˚ x 10.9˚ of visual angle) adapted from the NimStim stimulus set (Tottenham et al., 2009), varying in eye and mouth identity, were used to create a linear morph, yielding five stimuli varying in 25% increments. (Since the actual images used for experimentation are not publishable, all figures use example morphs from a different stimulus set.) All faces (Figure 2.1A) were cropped of external facial features using the same selection boundary shape (ellipse, 3 pixel feathering) and set to grayscale bitmaps in Adobe® Photoshop®.

The similarity of the resulting face images was analyzed using a biologically motivated, multi-scale, Gabor-filter model of V1 cortex (Renninger & Malik, 2004). A multi-dimensional scaling (MDS) analysis of the computational similarity scores revealed that, as expected, the faces varied along a single dimension and had roughly equal spacing between the 5 stimuli (spacing between adjacent, nominal 25% morphs: 30%, 24%, 21%, 25%).

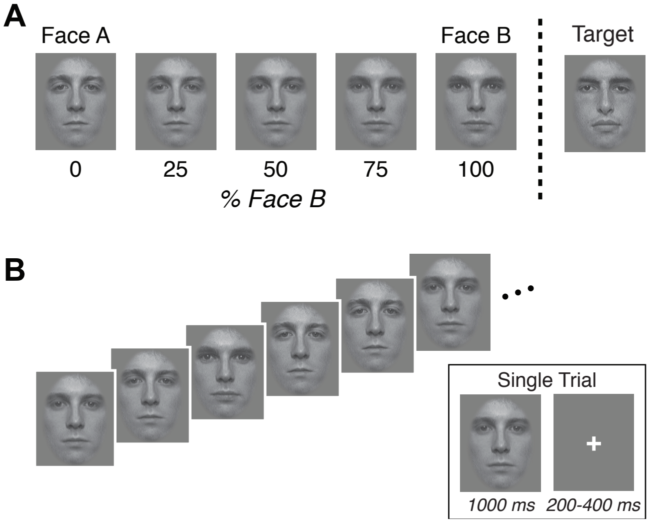


Figure 2.1: Example stimuli and presentation.
Representative example stimuli are presented here as the actual stimuli used were not approved for publication. (a) The experimental stimuli consisted of five faces morphed in identity between two endpoint identities (Face A and B) in 25% increments; subjects were not informed of the stimulus space arrangement. Subjects were instructed to monitor for the appearance of a target face (far right) whose identity was distinct from the morph axis. (b) Stimulus presentation. Stimuli were presented for 1000 ms with an ISI of 200, 300 or 400 ms, counterbalanced across trials using a type 1, index 1 sequence (Aguirre, 2007) with 18 elements.

2.3.3 Behavioral Assessment of Stimulus Similarity

A behavioral, reaction time study was used to confirm the monotonic ordering of the perceptual similarity of the stimuli along the face morph continuum. All subjects (N = 6) from the ERP study participated in the behavioral study several days following ERP data collection.

The 5 faces from the morph continuum were used as stimuli and presented side-by-side on a computer screen using the PsychToolbox (Brainard, 1997; Pelli, 1997) for MATLAB (Mathworks, Andover, MA). Subjects were instructed to respond with a button press to indicate if the pair of faces were the same or different (buttons indicating same or different were randomized to right or left across subjects). Each trial consisted of a side-by-side face presentation lasting until the subjects responded with a button press, followed by a 250 ms inter-trial interval. Runs consisted of 640 trials, with breaks occurring every 40 trials. "Same" trials, in which the face identity was the same, occurred with equal frequency as "different" trials. Within the "different" trials, the metric distances (Δ25, Δ50, Δ75, Δ100) along the morph continuum occurred with equal frequency.

For each different face pair for each subject, the inverse of the median of correct reaction times was found and entered into a distance matrix for multi-dimensional scaling (MDS) analysis (Kruskal & Wish, 1978). MDS analysis for each subject was performed for each subject using the MATLAB cmdscale() function. Coordinates were centered about the 50% face for each subject, and then averaged across subject to yield estimates of stimulus placement. The first dimension of the MDS estimate was retained.

2.3.4 ERP Stimulus Presentation

Each run consisted of 648 trials; each subject underwent 3 consecutive runs for a total of 1944 stimulus presentations. Each trial consisted of a stimulus presentation for 1000 ms, followed by an ISI of 200, 300, or 400 ms (counterbalanced across trials). Stimulus order was determined by a first-order, counter-balanced, n=18, type 1, index 1 sequence (Aguirre, 2007). An 18-element sequence was required to counterbalance the 6 stimuli (5 morphs and 1 target) crossed with the three durations of ISI that could follow each stimulus. During the ISI a central white fixation cross was presented on the same mean gray background surrounding the stimuli. Subjects were instructed to respond with a button press to the occurrence of a target face from outside the morph continuum (Figure 2.1A, far right). Subjects were trained on a simplified version of the task immediately prior to the experiment to ensure accurate identification of the target face. Target trials and trials immediately following target presentations were excluded from the main analyses.

Stimuli were presented using EPrime 2 (Psychology Software Tools, Inc.) on a Dell 24 inch LCD display situated 100 cm from the subject at eye level. Task responses were also collected through EPrime 2. To obtain “sensors of interest” for experimental analysis, after the main experiment subjects completed a short “localizer” experiment with faces, houses, and everyday objects (100 exemplars each), randomly interleaved. Stimuli in the localizer were presented on a white background with a black fixation cross (9.2° x 7.7° visual angle) for 300 ms (ITI jittered between 900 and 1100 ms); subjects were instructed to passively view the stimuli.

2.3.5 ERP Data Collection

Data collection was performed on a BioSemi ActiveTwo system (http://www.biosemi.com/products.htm) with 128 active electrodes with sintered Ag-AgCl tips in fitted headcaps. Evoked brain potentials were digitized continuously at a sampling rate of 512 Hz with default low-pass filtering at 1/5 of the sampling rate (http://www.biosemi.com/faq/adjust_samplerate.htm). Two additional electrodes with a 4mm sintered Ag-AgCl pallet were also placed bilaterally on the mastoids as references for data import (http://www.biosemi.com/faq/cms&drl.htm). Electrical offsets were verified to be between -20 and 20 µV for every channel prior to data collection.

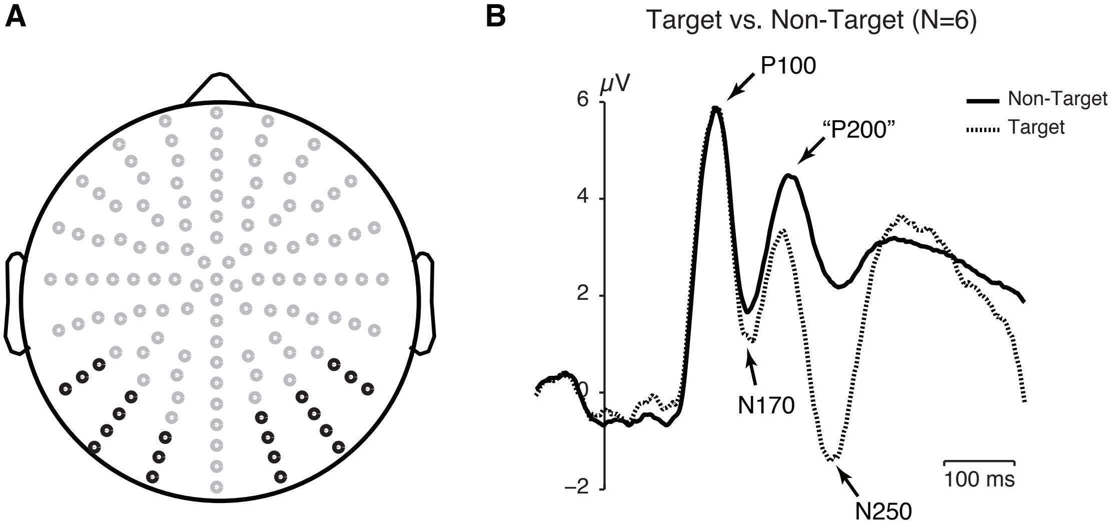


Figure 2.2: ERP sensor of interest (SOI) selection and component definition.
(a) Twenty-one face-selective (black dots) SOIs were selected across subjects using an independent localizer task (Face > House). (b) Component identification. Grand-average waveforms (N = 6) comparing the response to trials in which the target face was presented and all non-target trials. The P100 and N170 are the first positive and negative deflection, respectively. The N250 is functionally defined as having a greater negative deflection for target recognition (Tanaka et al., 2006).

2.3.6 ERP Pre-Processing and Analysis

Data were processed offline using the EEGLAB toolbox (Delorme & Makeig, 2004) for MATLAB. Sensors were selected for analysis using a "sensor of interest" (SOI) approach (Liu, Harris, & Kanwisher, 2002), via a point-to-point t-test comparing face and house conditions in the "localizer" scan. Significant channels for each subject were identified within the N170 and N250 latency ranges, and group channels (Figure 2.2A) used for subsequent analysis were selected if they were identified as significant in a majority of subjects (4 out of 6). Group average waveforms across all non-target trials for each sensor can be found in Supplementary Figure 1.

All data for each subject were saved from BioSemi ActiView and imported by run directly into EEGLAB. Mastoid channels were indicated as references to EEGLAB upon import and excluded; data were re-referenced immediately to the average signal of all 128 cranial channels. Data were epoched to a time window of 700 ms (100 ms pre-stimulus onset and 600 ms post) and baseline corrected (100 ms pre-stimulus onset). Trials containing artifacts (e.g., eye blinks) were identified and removed automatically using a ±100 μV threshold (average rejection rate across subjects for trials used in the main analysis was 16.7%, with a range of 5.3% - 38.8%).

ERP components of interest were identified for each subject individually using data averaged over all non-target conditions across the "sensors of interest" defined at the group level (Figure 2.2B). The previously-described P100 and N170 were defined on the basis of latency and direction of deflection, while the N250 was defined by the comparison of target and non-target faces (Tanaka et al., 2006). Inspection of our results also revealed a meaningful deflection between the N170 and N250, here called the P200. For each subject's grand average waveform, the time points of the local minima (for N170 and N250) and local maxima (for P100 and P200) were identified within search windows (P100: 125-175 ms; N170: 175-225 ms; P200: 225-275 ms; N250: 300-350 ms) and used as centers of the respective components for that subject. For each subject, the value of each component for each trial in each condition was then determined as a sum of the seven data points surrounding and including the subject's component center (approximating a 13.6 msec integral about the component center).

This area measure was computed for each trial, rather than across the trial-averaged data, to facilitate modeling of the data using a general linear model (GLM). Though commonly employed in fMRI analysis, GLM is rarely applied to ERPs. However, the GLM approach is methodologically superior for studies of similarity space, as it provides unbiased parameter estimates of both the "direct effect" (Aguirre, 2007) of morph identity, and of carry-over effects associated with similarity to the preceding face. If direct effects alone had been measured, the amplitude for (e.g.) the extreme Face A would be influenced by the tendency of that extreme Face A to be preceded by dissimilar faces, and thus be subject to less adaptation. Simultaneous estimation of the direct and carry-over effects in the context of a counterbalanced stimulus order allows the estimates to be efficient and unbiased. Similarly, as each condition in the non-geometric bias model represented a different subset of face identities, the simultaneous modeling of this effect and the direct effects ensures unbiased estimation of each.

For each subject, the data for each component (P100, N170, P200, N250) were entered into a general linear model composed of 11 covariates. Five covariates coded for the particular morph identity (Figure 2.1A) presented on any one trial: the "direct effect" of a given morph identity upon the amplitude of an ERP component. The remaining covariates modeled carry-over effects, or the effect of the status of the prior trial upon response amplitude for a given trial. Five of these covariates modeled the different sizes of change in stimulus identity between one trial and the next (Δ0%, Δ25%, Δ50%, Δ75%, Δ100%; Figure 2.4A); each covariate modeled those trials which had the given amount of identity change. A final covariate modeled asymmetric bias, and was set to have a positive value for trials in which the preceding trial was at the extreme of the morph continuum (0% or 100%) and the current trial at the center (50%), and a negative value for transitions in the other direction (from 50% to 0% or 100%). Trials in which the target face was presented, and the trials that followed target face presentations, were excluded. The estimates obtained from this first-order analysis were then collected across subjects into a second-order, random effects ANOVA analysis to test hypotheses of interest.

2.4 Results

In this experiment, we explored the time course of perceptual similarity by recording ERPs during face perception. Given that behavioral judgment of similarity has been hypothesized to consist of geometric effects of stimulus similarity and non-geometric effects of stimulus-specific bias, we tested if graded neural adaptation in the ERP data was consistent with this dual-effects model.

2.4.1 Behavioral Measure of Perceptual Similarity

To confirm that the stimuli were linearly ordered in perceived similarity, we collected a behavioral measure of similarity in all subjects. All subjects participated in a paired-discrimination task using the face stimuli. Accuracy across subjects was sufficient (mean d' 2.15) to allow an analysis of reaction time effects. An MDS analysis was conducted for each subject on the average reciprocal reaction time for each face pairing, and then averaged across subjects. Figure 2.3 presents the position of the five faces on the first MDS dimension, which accounts for 55% of the variance. As can be seen, the first dimension contained a monotonic ordering of the stimuli, with somewhat greater spacing of the faces away from the 50% morph. There was substantial agreement across subjects on the perceptual similarity of the stimuli as demonstrated by the small across-subject error bars. This ordering of the stimuli confirms that, as expected from the stimulus design, subjects perceived a monotonic perceptual change in identity across the face morph continuum.

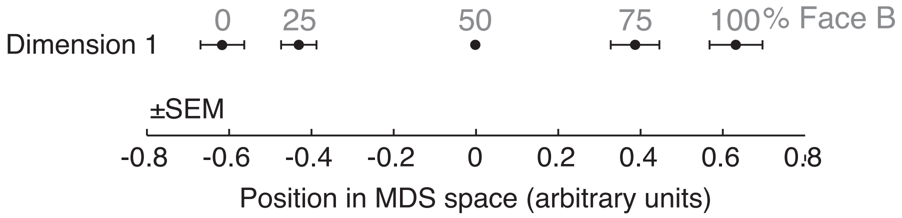


Figure 2.3: Behavioral results.
Inverse reaction times from a paired discrimination task from each of six subjects were entered into a multi-dimensional scaling analysis, with the resulting coordinates centered about the 50% face. The first dimension of the resulting model is displayed, which orders the faces monotonically along the morph continuum. This first dimension accounts for 55% of the variance. Error bars indicate plus/minus standard error of the mean across subjects.

2.4.2 Geometric Effect of Stimulus Similarity in ERP responses

ERP data were collected while subjects viewed a continuous stream of stimuli from the face continuum, presented in a counter-balanced order. ERP responses were assessed in relation to the identity of the face being presented, as well as the relationship of the current stimulus to the prior stimulus.

We first tested for a geometric effect of stimulus similarity based on the absolute metric distance from the preceding stimulus to the current stimulus along the face identity continuum. Data from each trial were binned depending on the morph distance between the face shown and the previous image, resulting in five similarity distances (Δ0, Δ25, Δ50, Δ75, Δ100). Thus, a distance of Δ0 would be a repetition of the identical stimulus, whereas Δ100 represented a stimulus at one extreme of the morph continuum following the face at the opposite extreme (Figure 2.4A).

Because of the monotonic ordering of the perceptual similarity space used here we would predict that the representation of metric stimulus similarity should change monotonically as a function of perceptual distance. In particular, given previous findings of neural adaptation in MEG (Furl et al., 2007; Harris & Nakayama, 2007; Harris & Nakayama, 2008) and ERP (Itier & Taylor, 2002; Jacques & Rossion, 2006; Kovács et al., 2006), we would predict greatest attenuation for Δ0, the identical repetition condition, with decreasing adaptation for increasing perceptual distances between stimuli.

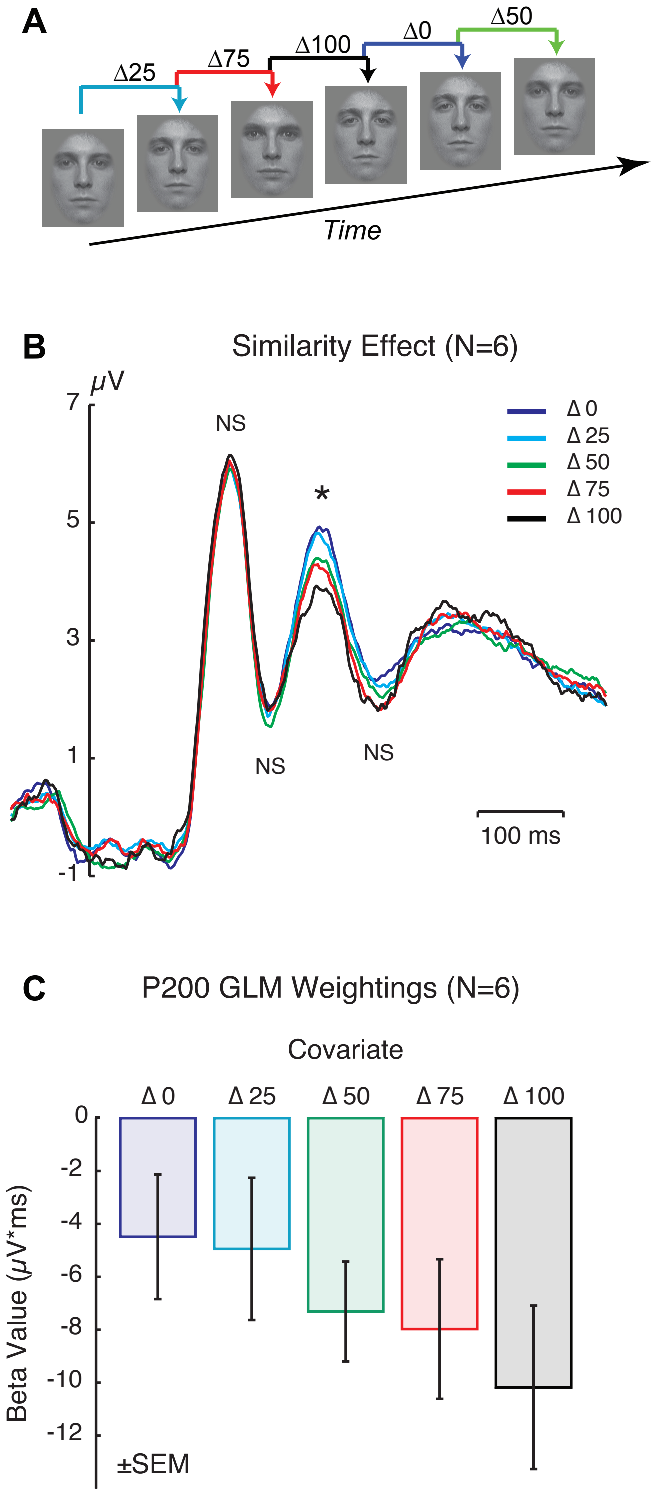


Figure 2.4: Geometric effect of similarity

(a) Trials were grouped based upon the metric distance of the preceding stimulus to the current stimulus along the morphed face continuum. Trials in which the target face was the current or preceding stimulus were excluded from analysis. (b) Grand-average waveforms (across all significant sensors; Figure 2.2) comparing each distance transition condition. A significant interaction of component and distance condition was observed, and within the P200 component there was a significant effect of distance (asterisk). Y-axis is aligned to stimulus onset. (c) Group average beta-values from the P200 for the five covariates modeling each distance condition in the general linear model. A significant effect of distance was observed, with a significant linear contrast. Error bars correspond to the between-subject SEM.

Grand average waveforms across all significant ERP channels (Figure 2.2) for each perceptual distance condition are displayed in Figure 2.4B. While the early perceptual P100 and N170 components showed no discernible effect of stimulus similarity, a graded adaptation effect is clearly visible between the N170 and N250 components. The most positive deflection for this component occurs in the Δ0 condition, with decreasing amplitudes for greater perceptual distances. Modulation of the P200 component, therefore, appears to index the earliest stage of processing associated with computations of metric stimulus similarity. Caution is required in interpreting these average plots, however. As discussed previously, apparent graded responses in the waveforms could result not from an adaptation effect, but instead from the unbalanced representation of particular face identities in a given dissimilarity pair (see Supplementary Table 1).

To test this finding in an unbiased manner, beta values from the general linear model were obtained for each subject and component, representing the weight on covariates modeling each absolute distance condition. These measures are independent of any "direct-effect" of stimulus identity (e.g., a hypothetically larger response to the extreme Face A or Face B). A repeated-measures ANOVA with component (P100, N170, P200, N250), and perceptual distance (Δ0, Δ25, Δ50, Δ75, Δ100) as factors showed a significant interaction between component and distance [F(12, 60) = 5.05, p = 0.00001], confirming that the effect of stimulus similarity is not seen for all components. Follow-up one-way repeated-measures ANOVAs for each component found a significant main effect of distance for the P200 [F(4, 20) = 6.01, p = 0.002] (Figure 2.4C), but no other components (all F tests < 2.8, *p*s > 0.05). The adaptation effect at the P200 was well-modeled by a linear contrast [F(1, 5) = 12.9, p = 0.016]. While a similar ordering of the adapted response is visible in the grand average waveform at the later N250 (Figure 2.4B), this effect was not significant (F(4, 20) = 3.38, p = 0.125).

Therefore, these data suggest that neural sensitivity to perceptual similarity begins within the first 400 ms of perceptual processing after stimulus onset. While the early perceptual P100 and N170 components do not show an effect of stimulus similarity, graded neural adaptation related to symmetric perceptual distance can be seen at the stage of processing following the N170, the P200 response. Along with its temporal position between the N170 and N250, this finding could be interpreted as placing the P200 at an intermediate cognitive stage between perceptual and mnemonic encoding.

2.4.3 Non-Geometric Effect of Asymmetric Bias in ERP Responses

In addition to the geometric representation of stimulus similarity, we also tested for non-geometric, asymmetric neural representation of the stimulus space. Given behavioral findings demonstrating a bias for more ‘prototypical’ stimuli (Op de Beeck, Wagemans, & Vogels, 2003), we hypothesized that the central face in the set, being an average of the faces at the extremes, would yield a differential effect on neural adaptation depending on whether it was a prior or current stimulus.

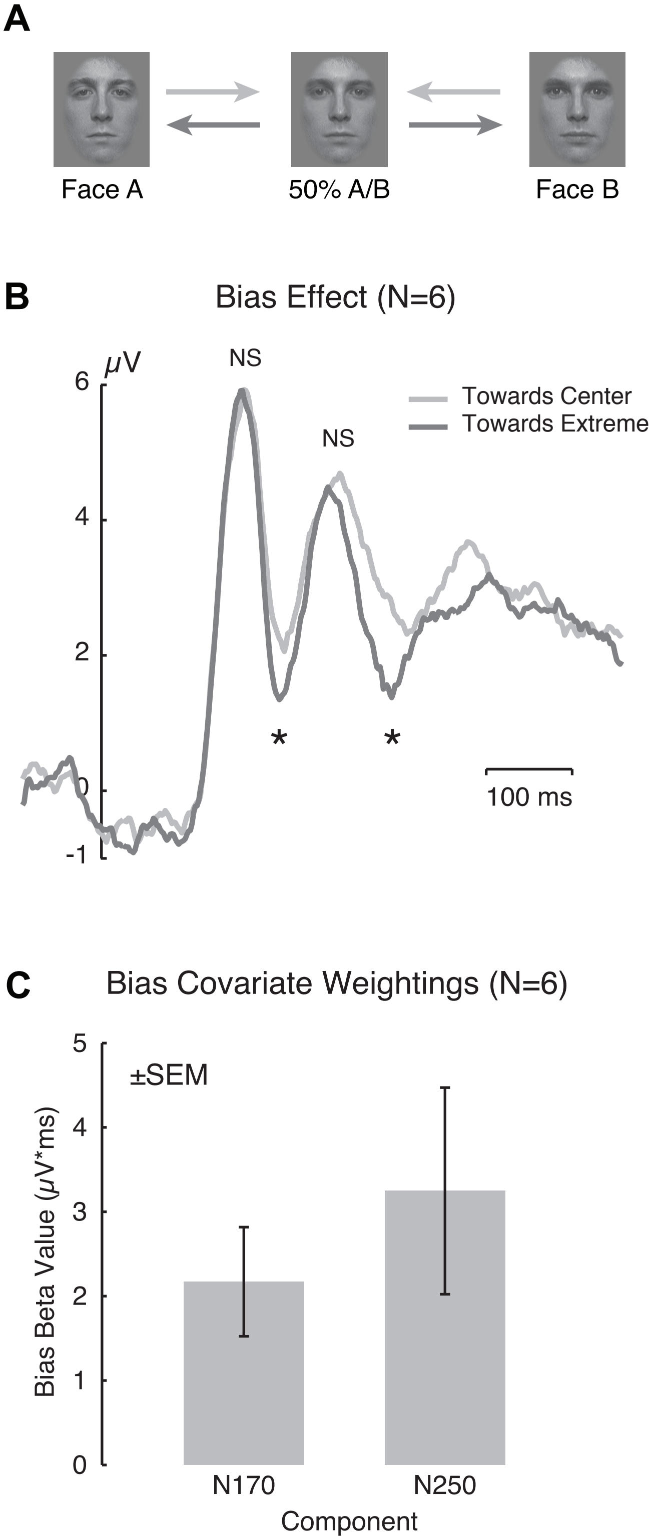


Figure 2.5: Non-geometric effects of similarity
(a) Trials were grouped based upon the direction of transition. “Towards center” trials were those in which the 50% face was presented following a face at either extreme of the morph continuum. “Towards extreme” trials had the opposite transition. (b) Grand-average waveforms (across all significant sensors; Figure 2.2) comparing each condition. A significant interaction of component and direction condition was observed, and significant effects of direction were observed within the N170 and N250 components (asterisks). Y-axis is aligned to stimulus onset. (c) Group-average beta-values from the N170 and N250 components for the covariate modeling the bias condition. Both components had significant weighting on the bias covariate, showing greater adaptation for the “towards center” transition in line with described prototype effects. These beta estimates are corrected for any “direct” effect of stimulus identity (i.e., a tendency for a larger amplitude response to “extreme” faces).

We compared the response on trials in which the central face is preceded by either of the two faces on the extreme of the stimulus space to trials in which the extreme faces are preceded by the central face (Figure 2.5A). Crucially, both of these conditions represent the same metric distance transition (Δ50), but vary in the direction of transition ('towards the center' of the stimulus space, and 'towards the extremes'). Previous work has proposed that extreme stimuli preceded by more central or prototypical stimuli are perceived as more dissimilar than central stimuli preceded by extremes (Tversky, 1977; Op de Beeck et al., 2003). Therefore, we predicted that neural adaptation would be sensitive to the direction of stimulus transition, with greater neural adaptation for transitions towards the center and less adaptation towards the extreme.

A group average of the two bias conditions is plotted in Figure 2.5B. In line with our predictions, transitions from the center of the stimulus space towards the extremes yield a greater negative deflection—but only at the N170 and N250 components. In contrast, the P100 and P200 display equal adaptation for both presentation orders. Again, these average waveforms confound direct and carry-over effects due to unbalanced representation of transitions and face identities (see Supplementary Table 2).

To evaluate the statistical significance of this effect, we modeled the stimulus transition as a covariate in a general linear model analysis. Loading on this covariate indexes the asymmetric carry-over effect of the transition, independent of other symmetric carry-over or direct effects. A repeated-measures ANOVA for the single bias covariate with component (P100, N170, P200, N250) as a factor showed a significant main effect of component [F(3,15) = 7.536, p = 0.003]. Follow-up one-sample t-tests across subjects indicated this asymmetric bias is significant in the N170 [t(5) = 3.36, p = 0.02] and N250 [t(5) = 2.65, p = 0.045] components (Figure 2.5C).

Thus, asymmetric bias effects also occur within the first several hundred milliseconds of visual processing. Interestingly, in contrast to the N170 and N250 responses, the P200 showed no significant asymmetric bias. This suggests, regardless of how geometric and non-geometric effects of similarity interact psychologically, the earliest neural stages associated with these computations are temporally separated. The visible asymmetric bias at the relatively early N170 response may be indicative that such bias effects need not rely on higher-level conceptual processing, but may be extracted relatively rapidly and early in the visual processing stream.

2.4.4 Direct Effects of Stimulus Identity on ERP Responses

Finally, we examined the “direct” effect of stimulus identity upon the ERP response. Studies of “prototype” responses in fMRI to faces, for example, have reported that there is a larger amplitude of neural response to distinct, as opposed to typical, stimuli (Loffler et al., 2005).

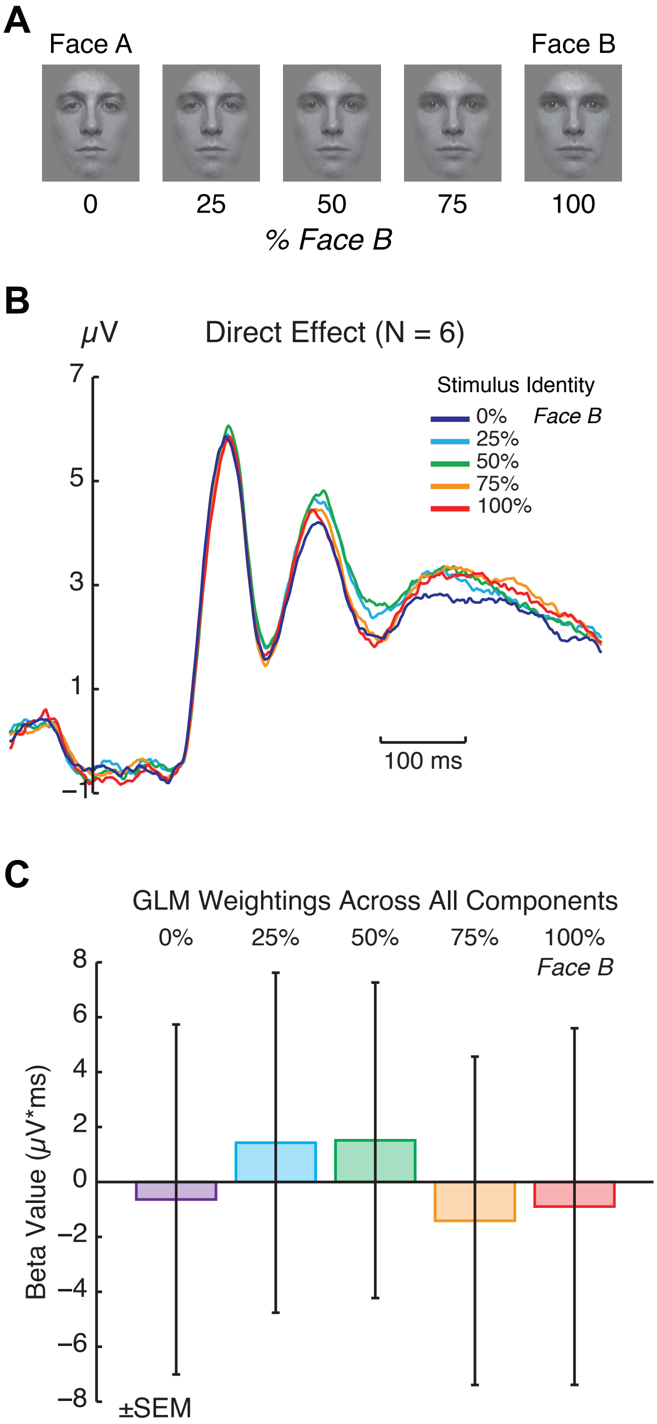


Figure 2.6: Direct effects of stimulus identity
(a) Trials were grouped based upon the identity along the morph continuum shown. (b) Grand-average waveforms (across all significant sensors; Figure 2.2) comparing each identity condition. A significant main effect of identity was observed, but no significant interaction of identity and component. Y-axis is aligned to stimulus-onset. (c) Group-average beta values collapsed across component are shown. As there was no significant main effect of component, or interaction of component with identity condition, beta-values were mean-centered within component for each subject, averaged across component for each subject, and then averaged across subject for display. Error bars correspond to between-subject SEM of mean-centered, across-component averages.

A group average of the stimulus identity conditions is presented in Figure 2.6B. Some separation between the identities is visible in the P200 and N250 components, perhaps consistent with a differential response to the extreme stimuli from the morph continuum as compared to the center. As discussed previously, however, these effects may be confounded by carry-over effects. For instance, a grand average waveform for the "direct" effect of the 50% morph face is confounded by the fact that the 50% morph is, on average, more often preceded by similar faces by virtue of its central location; and thus more subject to adaptation. Similarly, a postulated differential response to the 50% morph face compared to the extreme faces (a "direct" effect) might confound the non-geometric bias effects without concurrent modeling.

To examine direct effects in an unbiased manner, we obtained the beta values associated with the amplitude of the ERP response to each face identity, after accounting for the adaptation and bias effects. A repeated-measures ANOVA was then performed with each identity covariate (0%, 25%, 50%, 75%, 100% Face B identities) and component (P100, N170, P200, N250) as factors. A significant main effect of identity was found [F(4, 20) = 5.444, p = 0.004], but the interaction of identity and component was nonsignificant [F(12, 60) = 1.400, p = 0.191], suggesting this main effect of identity did not differentially modulate any component in particular. Figure 2.6C presents the average across subjects and components of the response to each face identity. The pattern of responses does not correspond readily to a simple model of prototype or geometric effects.

2.5 Discussion

Psychological models of perceptual proximity, the subjective judgment of “likeness” between stimuli, have historically drawn a distinction between two factors or processes: representation of simple metric distance between stimuli, and stimulus-specific bias. Quantified in models such as the ‘additive similarity and bias’ model (Holman, 1979; Nosofsky, 1991), this two-part framework separating geometric and non-geometric effects has guided our understanding of how the visual system represents variation between stimuli.

What are the neural correlates of these processes? We examined this question using a continuous carry-over design (Aguirre, 2007) in ERP. Previously used in fMRI, continuous carry-over designs allow measurement of graded neural adaptation, and therefore better characterization of the neural representation of perceptual similarity space. Using this paradigm with a set of ordered, morphed faces in ERP, we tested a dual-effects model of perceptual similarity against a temporal framework of early visual evoked components previously associated with face processing.

Modeling transitions between stimulus presentations in terms of absolute metric distance along our morphed face continuum, we found graded neural adaptation consistent with metric stimulus similarity at a component between the N170 and N250 responses. Modulation of the P200 was related to perceptual similarity, with greater positive deflection for smaller perceptual distances (Figure 2.4). The temporal position of this component suggests that computation of metric stimulus similarity begins within the first several hundred milliseconds of stimulus presentation, although after the earliest stages of perceptual processing indexed by the P100 and N170 components. Adaptation of a neuroimaging signal that is proportional to stimulus similarity can result from a cortical region that codes stimulus identity by a population code (Aguirre, 2007; Drucker, Kerr, & Aguirre, 2009). This suggests that, at the P200 stage, a neural population code for facial identity is evoked that reflects geometric effects of similarity. It is also possible that another neural mechanism apart from adaptation (e.g., a re-entrant masking effect; Kotsoni, Csibra, Mareschal, & Johnson, 2007) is responsible for this parametric modulation. In either case, these data are among the first to place a neural signature of geometric similarity coding within a definite time window, arising as early as 200 ms after stimulus presentation.

We also modeled the effects of asymmetric bias (Tversky, 1977; Op de Beeck et al., 2003). Neural markers of such a non-geometric similarity effect were found for the N170 and N250 components (Figure 2.5). While both the N170 and N250 components show sensitivity to asymmetric transitions positioned about the center of the stimulus space, the P200 does not. Thus, not only have we found neural correlates of perceptual proximity processing within relatively early stages of perceptual processing, but we also demonstrate that the encoding of metric stimulus similarity and asymmetric bias are temporally distinct.

Our model of non-geometric similarity effects is based upon the notion of a ‘prototype’ effect (Tversky, 1977; Op de Beeck et al., 2003). Two stimuli are perceived as more proximal when the more prototypical or average stimulus is presented following another one less so, and less proximal in the reverse case. There are other non-geometric bias effects that might be considered. In studies of magnitude estimation, for example, the response to a stimulus tends to be larger when the preceding stimulus intensity was greater. This “assimilation” effect is commonly seen for stimuli in which one end of the continuum is “larger” (DeCarlo & Cross, 1990). The opposite, “contrast” effect is also observed. A model for this directional bias in neuroimaging data is considered in Aguirre (2007), and is orthogonal to the ‘prototype’ effect just discussed. While the ‘prototype’ model of bias is symmetric about the center of the stimulus space, directional bias is inversely symmetric towards each extreme. Directional bias has been observed in perceptual adaptation effects for face identity (Leopold, O'Toole, Vetter, & Blanz, 2001), gender (Webster, Kaping, Mizokami, & Duhamel, 2004), and attractiveness (Rhodes, Jeffery, Watson, Clifford, & Nakayama, 2003). We tested for directional bias effects in our ERP study but found no significant effect (data not shown). This is not surprising as our stimuli were a morph between two faces of equal distinctiveness, as opposed to the stimuli of (e.g.) Leopold et al. (2001) in which one end of the continuum was a distinctive face and the other a prototypical or average face.

A perceptual prototype may arise from long-term exposure to stimuli of a given class, from the local context of a set of stimuli in an experiment, or both. Our study did not distinguish between these two types of prototype. The center point of our stimulus continuum may have achieved prototype status as it was a more “average” face in general, or because it was the central tendency of this particular stimulus set. These possibilities might be distinguished through the use of an unbalanced face continuum in which the "middle" face in the local context of the experimental set is not the most average at the global level.

Related to this point, it is worth noting that while we observed neural prototype effects for both the N170 and N250 components, it is possible that these distinct components are related to different prototype effects. For the N170 in particular, we might expect that the "prototype" effect reflects a local stimulus effect, driven by the experimental stimulus space alone. Previous work has demonstrated a lack of adaptation in the N170 to within-class features of faces, including eye-gaze direction (Schweinberger, Kloth, & Jenkins, 2007) and gender (Kloth, Schweinberger, & Kovács, 2009). These findings suggest that the N170 adapts in a broad categorical fashion to faces and not to within-category features, such as global face distinctiveness. Taken together with the apparent role of the N170 in structural encoding (Bentin et al., 1996; Rossion et al., 2000), we would suggest that the "prototype" effect observed in the N170 might reflect a rapid, implicit extraction of local central tendency (i.e., within the experimental stimulus space). In contrast, as the N250 is thought to reflect access to stored face representations (Tanaka et al., 2006), it is possible that the non-geometric effect observed in this component indexes transitions about a stored, global face "prototype". While these interpretations rely on the characteristics of the underlying components, future experiments which dissociate local and global face prototypes in the manner described above could characterize putatively separable non-geometric similarity effects in a component-independent manner.

Finally, a notable methodological feature of this study was the concurrent measurement and separation of the direct effects of each stimulus from carry-over effects of adaptation and asymmetric bias. Without explicit modeling, these effects are confounded, rendering it unclear whether effects reflect perceptual proximity per se, or a combination of adaptation and identity effects. This potential confound exists in several studies of face representation. For example, Loffler et al. (2005) used a block design in fMRI to demonstrate increasing BOLD signal in the fusiform face area (FFA) in response to groups of faces of increasing ‘distinctiveness’. The authors define ‘distinctiveness’ as distance along putatively orthogonal identity axes extending from a central ‘mean’ face. This design focuses primarily on non-geometric prototype and identity effects. However, their observed decrease in BOLD signal for face blocks more proximal to the mean could represent neural adaptation indexing geometric effects of metric distance, or some combination of geometric and non-geometric effects.

Likewise, in an fMRI study using a similar facial identity morph continuum to ours, Jiang et al. (2006) reported non-linear BOLD adaptation in response to increasing metric distance. The authors interpreted this finding as suggesting that neural adaptation would asymptote for greater metric stimulus distances, something we do not observe in our data. In their experimental design, Jiang et al.(2006) use a traditional paired-presentation paradigm with the adapting stimuli only located at the extreme of the morph continuum, and test stimuli at Δ30, Δ60 and Δ90 metric distances. It is possible with this design that the unbalanced frequency of stimulus presentation introduces a non-geometric similarity effect such as the ‘relative prominence’ bias presented by Johannesson (2000), or an asymmetry driven by exposure frequency as presented by Polk et al. (2002). Thus while Jiang et al. (2006) suggest their data reflects non-linear (asymptotic) encoding of metric linear distance, our findings suggest their data could reflect a combination of geometric effects and non-geometric effects.

2.6 Conclusions

Our results provide evidence for the dissociation in neural coding of non-geometric ‘prototype’ effects from the geometric effects of stimulus similarity, supporting psychological models of the two elements as separate factors in the perception of proximity. Using a continuous carry-over design in ERP, in conjunction with a principled GLM approach to distinguish geometric and non-geometric processing, we find that these different effects occur at discrete temporal stages of face processing. These findings should expand our understanding of neural similarity, offer new avenues for exploring global and local prototype effects, and encourage more careful consideration of the complexity of stimulus space representations in the brain.

3 Confounding of Norm-Based and Adaptation Effects in Brain Responses

3.1 Abstract

Separate neuroscience experiments have examined two properties of neural coding for perceptual stimuli. *Adaptation* studies seek a graded recovery from neural adaptation with ever greater dissimilarity between pairs of stimuli. Studies of *prototype* effects test for a larger absolute response to a stimulus which is distant from the center of a stimulus space. While intellectually distinct, these effects are confounded in measurement in standard neuroscience paradigms and can be mistaken for one another. Stimuli which are more distinctive are less subject to adaptation from perceptual neighbors. Therefore, a putative prototype effect may simply result from greater adaptation of prototypical stimuli by other stimuli in the experiment. Conversely, stimulus pairs which are the most perceptually distant from one another, and therefore expected to show the greatest recovery from adaptation, disproportionately draw from the extremes of the stimulus space. Thus, a putative neural similarity effect may be created by an underlying prototype representation. We simulate BOLD fMRI results driven by each possible effect and demonstrate spurious results in support of the complementary effect. We then present an example fMRI experiment that demonstrates the confound and how it may be minimized. Finally, we discuss the implications of this intrinsic confound for studies of perceptual representation, neural coding, and category learning.

3.2 Introduction

A common target of neuroscience studies is the form of neural coding used to represent variation in stimulus properties. Very often, such studies use stimuli with linear variation along a single dimension. Examples of these "morphed" stimuli include facial image morphs of identity (Freeman et al., 2010, Jiang et al., 2006, Kahn et al., 2010) or emotional expression (Said et al., 2010a), mathematically defined abstract shapes (Panis et al., 2010, De Baene & Vogels, 2010), or auditory cues (Latinus et al., 2011).

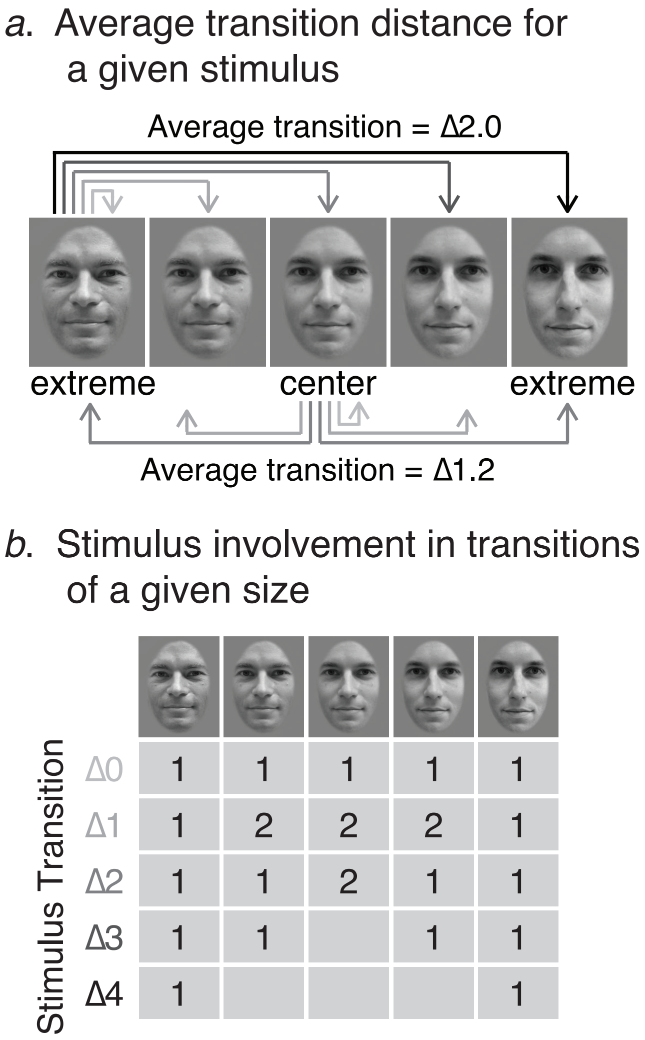
Within the broad category of distributed neural encoding models (Barlow, 1972, Edelman, 1998), perceptual variation can be expected to have several neural correlates. Norm-based encoding models (Leopold et al., 2001, Rhodes and Jeffery, 2006) postulate that variation relative to a reference point in a stimulus space results in differential absolute responses to stimuli. These differences may take the form of a "prototype" effect (Valentine, 1991): a reduction in the neural response to a centrally-oriented prototype relative to those stimuli that are more extremely positioned. However, other distributed encoding models are possible, including those in which a stimulus space is represented using tuning functions that do not depend upon a particular point of space as a reference. In such a case, as in all distributed encoding models, perceptual variation could be indexed by the overlap in neural populations constituting two distributed representations. One manifestation of this form of neural representation is an "adaptation" effect (Grill-Spector & Malach, 2001; Henson & Rugg, 2003): a reduction in the neural response to a stimulus resulting from recent presentation of an identical or related stimulus. As defined, these two effects of encoding are intellectually distinct and based upon related and well-defined schema for neural representation.

Testing for these two effects of perceptual variation is possible via neuroimaging. Prototype effects, hypothesized to manifest as a larger bulk neural response to extreme stimuli, have been observed using functional magnetic resonance imaging (fMRI) in response to faces (Freeman et al., 2010; Loffler et al., 2005; Said et al., 2010a), face profile silhouettes (Davidenko et al., 2011) and abstract shapes (Panis et al., 2010). Similar findings have been demonstrated in monkey electrophysiological recordings (Leopold et al., 2006). Adaptation effects, a form of "carry-over" effect of one stimulus upon another (Aguirre, 2007), manifest as an increasing reduction in neural response for the latter stimulus in a sequentially-presented pair as a function of the pair's dissimilarity in fMRI (Drucker et al., 2009, Jiang et al., 2006) and ERP (Kahn et al., 2010). Graded neural adaptation related to stimulus similarity has been demonstrated in MEG (Furl et al., 2007) and in neuronal firing in monkey electrophysiology studies (De Baene & Vogels, 2010).

Despite being coherent and distinct predictions of neural models, we show here that these effects are confounded in measurement, and thus can be mistaken for one another. Importantly, while counter-balance (Aguirre et al., 2011) in the order of stimulus presentation is ultimately necessary to address this confound, it is not sufficient to remove it.

3.3 An Example Stimulus Space

Consider a simple experiment that presents stimuli in a counter-balanced order from a set of five, evenly spaced morphed faces (Figure 3.1A; morphs created using Photoshop CS5.5, Adobe; & JPsychoMorph). We may then ask if different face morphs have systematically different relationships to the set of stimuli as a whole.



a. Average transition distance for a given stimulus

b. Stimulus involvement in transitions of a given size

Δ0	1	1	1	1	1
Δ1	1	2	2	2	1
Δ2	1	1	2	1	1
Δ3	1	1		1	1
Δ4	1				1

Figure 3.1: Consequences of a counterbalanced experimental design with a 5-exemplar morph space
(a) An example stimulus set consisting of the two authors of this paper morphed in 5 equal steps. The average distance of all possible transitions from the central face is less than that from either extreme face. (b) Relative representation of each stimulus in every possible transition distance for a counterbalanced stimulus sequence. Transition distance is measured as the number of steps within the stimulus space between the preceding and current stimuli.

Faces from the center of the space will, on average, be preceded and followed by faces which are more similar: on average, there will be a transition of 1.2 positions within the stimulus space from a center stimulus to the prior or next stimulus in the sequence (Figure 3.1A). In contrast, faces from the ends of the stimulus space will have transition sizes of 2.0 positions from sequentially adjacent trials on average. Thus, the position of the stimulus within the space is related to the size of transitions in which it is involved. If different neural responses attended stimulus transitions of different sizes, this relationship would produce different average neural amplitudes to the different faces, *even if the neural responses to the faces themselves were identical*. This is a mechanism by which neural adaptation to stimulus similarity alone might be mistaken for a prototype effect. This can be appreciated in the complementary analysis as well (Figure 3.1B). Consider the sizes of transitions that are available between the faces in the experiment. Only stimuli from the ends of the space can be involved in the largest transitions. Conversely, small transitions disproportionately involve the faces from the center of the space. If the faces from the ends of the stimulus space evoked larger neural responses than faces from the center, this relationship would produce different average neural responses to the transitions of different sizes, *even if there was no effect of transition size itself upon neural response*. This is a mechanism by which prototype effects alone might be mistaken for neural adaptation to stimulus similarity.

We note that the first of these concerns has been recognized previously (Panis et al., 2010; Davidenko et al., 2011). We expand upon these previous observations by highlighting the reciprocal nature of this confound (which affects more than just studies of norm-based encoding), describing steps to mitigate the problem, and illustrating the explanatory potential when this complexity is embraced by experimental designs rather than eliminated.

3.4 A Simulated Experiment

We conducted a simulation of an experiment that uses a linear morph space. Following the parameters of a recent study of prototype representation (Panis et al., 2010), we created a sequence for presentation of five stimuli (along with blank trials) using OptSeq2 (NMR Center; Massachusetts General Hospital, Boston, MA).[1] An inter-trial-interval of 2000 msecs was assumed.

We first simulated the case in which a neural population has norm-based (prototype) coding for the stimuli, but no neural adaptation takes place. Figure 3.2A (top row) shows the "carry-over matrix" (Aguirre, 2007) which characterizes the neural response to a given stimulus as a function of the prior stimulus. As can be seen, the modeled neural response is entirely determined by the identity of the current stimulus ("direct" effects). The particular amplitudes of response used were taken from the measure of a behavioral prototype effect (Upper left panel of Figure 4 of Panis et al., 2010).

[1] While the optseq program offers "preoptimized first-order counterbalancing", it does not actually provide perfect counter-balance of the stimuli (Aguirre et al., 2011). This has no consequence for the didactic purpose of our simulation, but would complicate attempts to remedy the confound within a linear model.

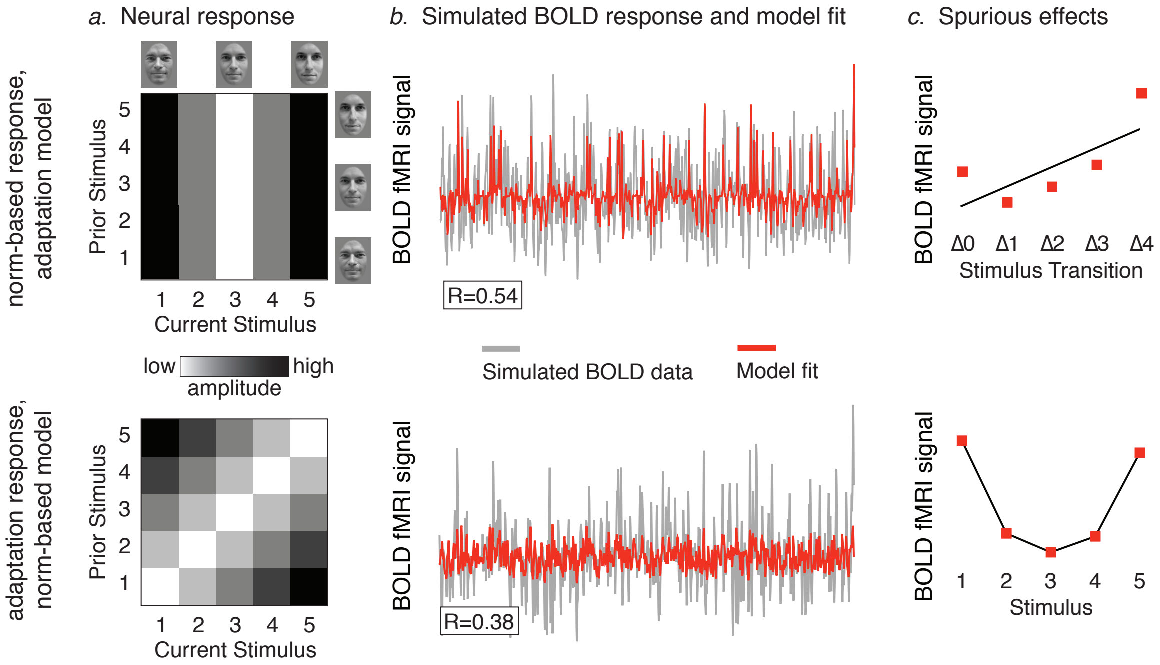


Figure 3.2: A simulation of fMRI BOLD response demonstrating confounding of prototype and similarity effects.

TOP ROW (a) A neural response model in which only prototype effects are postulated, with extreme stimuli resulting in a greater bulk response. The amplitude of neural response is driven entirely by the current stimulus with no modulation by preceding stimuli. (b) Simulated BOLD response for a counterbalanced stimulus presentation driven by prototype effects (grey). The model fit (red) represents a covariate modeling a linear adaptation effect for transition distance but not the prototype effect driving the data**.** (c) A spurious linear neural adaptive effect of similarity resulting from solely un-modeled prototype effects. BOTTOM ROW (a) A neural response model in which only stimulus similarity effects are present, with large transitions resulting in the greatest neural response (recovery-from-adaptation) and repetitions yielding the smallest. Individual responses are a function of the distance of the prior stimulus to the current stimulus (b) Simulated BOLD response for a counterbalanced stimulus presentation driven by similarity effects (grey). The model fit (red) represents a covariate modeling a prototype effect for transition distance but not the similarity effect driving the data**.** (c) A spurious effect of prototype resulting from solely un-modeled neural adaptive effects.

Given the sequence of stimuli and the matrix of neural responses, a simulated BOLD fMRI signal was generated (grey line, top row, Figure 3.2B) using an assumed hemodynamic response function (Aguirre et al., 1998).

We then analyzed the simulated data using a model that tested only for the presence of neural adaptation effects, and ignored the possibility of a prototype effect. In the model, covariates were generated to model transitions between the stimuli of different step sizes. As can be seen, the model of a non-existent neural effect fit a substantial portion of variance in the simulated BOLD data (red line, top row, Figure 3.2B). A plot of the loading on the model covariates reveals a spurious effect that could easily be mistaken for a linear neural adaptation effect (top row, Figure 3.2C). Therefore, in data that contain only "prototype" neural effects, a spurious neural adaptation effect might be found.

Next, we simulated the case in which a neural population scales the amplitude of response dependent upon the similarity of the prior stimulus in the sequence, but which has equal responses to all the stimuli in isolation (Figure 3.2A, bottom row). Again, simulated BOLD fMRI data were generated. These data were then modeled assuming that only direct effects of the stimuli are present in the data, and ignoring any possible neural adaptation. Separate covariates were fit to the average neural response to each stimulus identity (Figure 3.2C, bottom row). The result (Figure 3.2C, bottom row) is a spurious "prototype" effect, in which larger amplitude neural responses are measured for the stimuli from the extremes of the stimulus range. Therefore, in data that contain only neural adaptation effects, a spurious "prototype" effect may be measured.

3.5 An Empirical Example

We next collected fMRI data from one participant (naive to the hypotheses of this study) to demonstrate these spurious effects in practice, and mitigation of the confound through concurrent modeling of both effects. Stimuli were 5 radial frequency contours (RFCs; Op de Beeck et al., 2001) created along a linear axis of varying RFC-phase and amplitude, and rendered with a pseudorandom black checkerboard texture on a gray background. The stimuli, subtending 5˚ x 5˚ of visual angle, were back projected onto a screen and viewed by the subject via a head coil mounted mirror. The five stimuli, an additional target stimulus (an RFC orthogonally related to the morphed stimuli), and a blank trial were presented in sequences defined by second-order counterbalanced k=7, n=3, de Bruijn cycles (Aguirre et al., 2011). Each of 4 runs consisted of 343 continuous trials. Trials consisted of a 900 msec stimulus presentation followed by a 200 msec inter-stimulus interval of a grey blank screen (Figure 3.3A). The duration of all blank trials was either doubled or tripled pseudorandomly to fit the 343 trials to 154 TRs. The subject was directed to monitor for the appearance of the target stimulus and respond with a button press. Echo-planar BOLD fMR images were collected (TR = 3 sec), with 3 mm isotropic voxels on a Siemens 3-T Trio with a 8-channel head coil. A functional localizer (Harris & Aguirre, 2008) consisting of faces, objects, buildings, and phase-scrambled images was also run for use in region-of-interest (ROI) definition. We defined an ROI corresponding to the left ventral lateral occipital complex (LOC) that had a significantly greater response to objects and faces (as compared to buildings and scrambled images) and a significant response to the average (main effect) of all shape stimuli in the primary experiment as compared to a blank screen (Figure 3.3B). The response to the different RFC shapes and their transitions were obtained within this ROI.

The raw data were sinc-interpolated in time to correct for slice acquisition order and motion corrected using least squares minimization. The effects of adaptation and prototype in the data, both in isolation and concurrently, were analyzed using a modified general linear model (Worsley & Friston, 1995). After accounting for serial correlation in the residuals and the covariates used, the statistical tests we report below had approximately 110 effective degrees of freedom.

Our first model contained covariates only for adaptation, without differences in absolute response to individual stimuli modeled. Four covariates modeled the possible step sizes from the prior stimulus to the current stimulus (Δ1 through Δ4, as in Figure 3.1; identical stimulus repetitions, Δ0, served as a reference condition for the entire model to avoid over-fitting of the model to the data). Additional covariates modeled the main effect of stimulus presentation versus the blank trials, targets, transitions from blanks to a stimulus, and the whole-brain global signal. The weightings on these covariates are presented in Figure 3.3C, top-left panel. A significant effect of step size, (evaluated simply as a t-test for [Step 4 - Step 1]; $t = 3.46$, $p = 0.0004$) is present in this analysis, but is subject to potential confound of norm-based effects.

A complementary analysis modeled only the “direct” effect of each stimulus. Four covariates were fit the response to the presentation of each non-target stimulus,

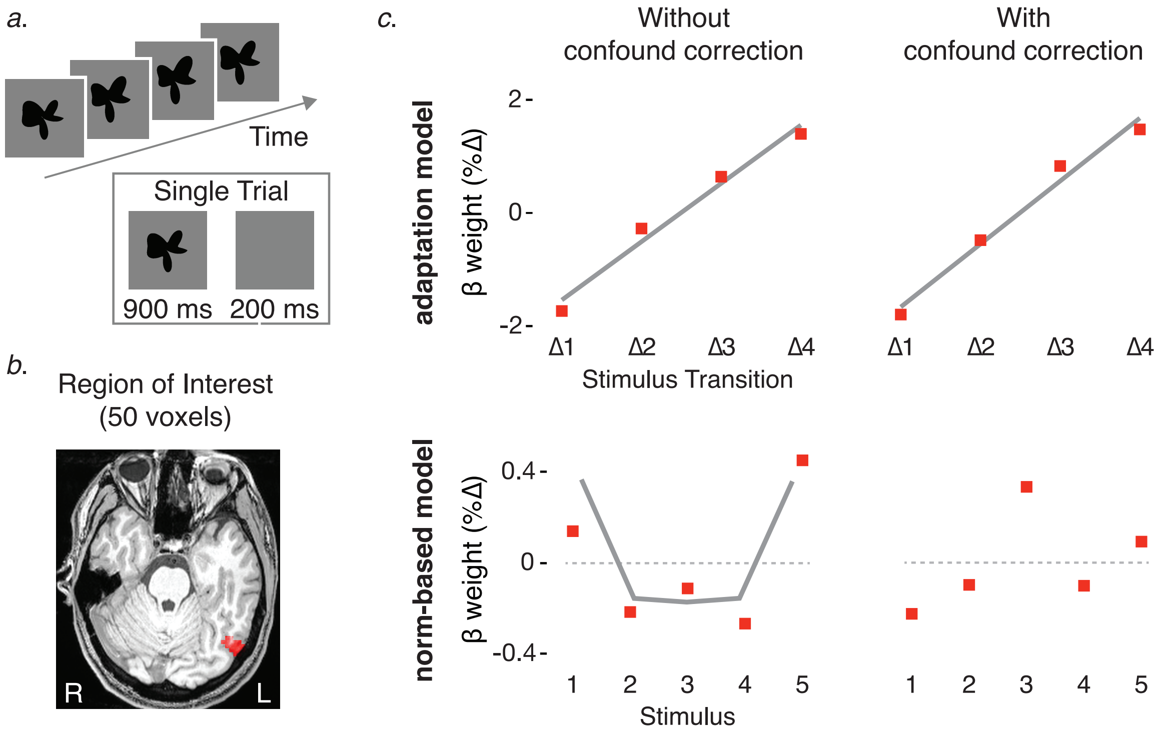


Figure 3.3: An analysis of empirical fMRI data demonstrating both confounded and unconfounded measurement of adaptation and prototype effects

(a) Experimental design. Individual trials consisted of a stimulus presentation of 900 ms followed by a blank screen ISI of 200 ms. Trials proceeded continuously while the subject monitored for the appearance of an unrelated target shape (not shown). (b) The region of interest (ROI) used for statistical analysis, 50 voxels in ventral LOC, defined using an independent localizer comparison of [(Faces, Objects) - (Buildings & Scrambled Images), t > 4] crossed with a main effect of [Shapes, t > 4]. (c) TOP ROW Beta values from general linear models for covariates modeling the adaptive effects of transition distance. All values are mean-centered, One the left, the effects of stimulus identity are un-modeled in the GLM. On the right, the effects of both adaptation and identity are modeled concurrently. With concurrent modeling, the strength of the adaptive effect can be quantified without confound. BOTTOM ROW Beta values from a GLM for covariates modeling stimulus identity, indexed to the central stimulus. One the left, the effects of adaptation are un-modeled in the GLM. On the right, both effects are modeled concurrently. With concurrent modeling, the trend of a prototype effect is no longer present.

referenced to the central stimulus. The additional covariates included in the model were the same as that for the prior, “adaptation only” analysis. The covariate weights for this model are presented in Figure 3.3C, bottom-left panel. The presence of a norm-based effect of prototype should manifest as a greater response for the extreme stimuli relative to the central stimulus. A t-test for [(stimulus 1 - stimulus 3) + (stimulus 5 - stimulus 3)] in this one subject showed an effect in this direction (t = 1.18, p = 0.12). Thus a norm-based effect of prototype could be present, but is similarly subject to confound due to un-modeled effects of adaptation.

A third model contained both sets of covariates. The resulting beta values are presented in Figure 3.3C, right panels. When controlling for the effect of prototype, the carry-over effects of adaptation observed in previous GLM remain in the larger model (evaluated as before, t = 3.59, p = 0.0002). However, when these carry-over effects are modeled in parallel, the suggestive trend of a norm-based effect of prototype, with the extreme stimuli yielding greater response than the central stimulus, is no longer present (t = -1.17, p = 0.87).

The ventral region we examined is close to that previously reported to manifest proportional adaptation for two-dimensional closed contours (Drucker & Aguirre, 2009), and which is not thought to demonstrate significant norm-based effects of prototype (Panis *et al.*, 2010).

With these data, we present an empirical example of the confound of adaptation effects and norm-based effects. We demonstrate that in the same data, incomplete modeling can lead to spurious effects for which complete modeling can account.

3.6 Implications For Other Studies

We simulated and measured this confound in a particular experimental design that presented the stimuli in a counter-balanced order, but it is present in other studies as well. In an fMRI study, Jiang et al. (2006) argued in favor of a non-linear trend in recovery from adaptation as a function of stimulus dissimilarity. The authors presented a series of facial stimulus pairs of varying inter-stimulus distances drawn from a linear morph space, and observed the predicted recovery-from-adaptation. However, as the authors used only an extremely-positioned stimulus as the adapting stimulus, and an uneven selection of test stimuli, it is possible that their measures of recovery-from-adaptation were confounded by un-modeled effects of prototypicality.

In another fMRI study, Loffler et al. (2005) presented blocks of faces varying in distinctiveness from an average face. The authors demonstrated an increase in neural response for blocks of faces further from the average face, a finding which was presented in support of a mean-centric direct effect. While the prototype explanation is possible, the experimental design is confounded in that faces more similar to a prototype are geometrically less distinct. Blocks of more prototypical faces would thus be subject to greater neural adaptation, yielding a reduction in response amplitude driven by carry-over effects. Davideko et al. (2011) subsequently used an elegant stimulus manipulation to call this result into question. Within blocks of stimuli, they manipulated the distinctiveness of parameterized face silhouettes while controlling the physical variability of the stimuli in the block. While this stimulus manipulation removes the confound, it does not provide a generalized solution to the joint examination of distributed and norm-based neural coding.

A similar stimulus set was used by Leopold et al. (2006) during electrophysiological recordings of inferotemporal cortex in monkeys. The authors demonstrated increased neural activity for faces further from the average face in support of a norm-based encoding model. While it is difficult to assess the potential for confound, (indeed, different experimental methods can minimize this potential, as we will discuss) this study demonstrates that stimulus sets vulnerable to confounding of prototype and adaptation are not limited to human neuroimaging.

The confound of prototype and adaptation effects will have a more subtle effect in multi-voxel pattern analysis (MVPA) studies that make use of one-dimensional stimulus sets (Said et al., 2010b). In this case, the uncertainty regards the form of neural representation that is used to decode the stimuli. MVPA studies make use of the pattern of direct-effects across voxels (the average response to a given stimulus across presentations). The ability of an MVPA analysis to classify the identity of a stimulus from the pattern of activity may not be the consequence of a difference in the neural response to the stimulus itself, but instead as a consequence of differences in relative neural adaptation.

Finally, a confound between norm-based and adaptation measures has implications for studies of category formation as well. A common hypothesis predicts enhanced recovery-from-adaptation for a stimulus transition of a given distance that crosses a perceptual category boundary, relative to a transition of the same distance that does not cross a category boundary (Goldstone, 1994). However, as category boundaries typically bisect a stimulus space (and large stimulus transitions crossing the boundary have no within-category analog), stimulus transitions crossing the category boundary will preferentially sample stimuli from the center of the space, while within-category transitions will involve more extremely positioned stimuli. In such a case, a smaller bulk response to centrally-oriented stimuli due to norm-based coding effects could negate or even *reverse* the predicted alteration in adaptation effects driven by a category boundary. Indeed, we are aware of results from our lab (and others) that demonstrate this reversal and have to date remained unpublished due to puzzlement regarding the cause and interpretation.

It is important to emphasize that the acknowledgment of this confound does not negate the presence of a neural effect in the studies we cite. Instead, this confound leads to uncertainty regarding the precise form of neural coding that produced a measured neural response.

3.7 Mitigation of Confound

With an understanding of the potential of confound between adaptation and direct effects, we can consider several steps that may be taken to mitigate the problem. It should be noted that the study we used as a model for our simulation (Panis et al., 2010) is also a model example of awareness of this possible confound. The authors considered the possibility of adaptation effects in their data, and conducted an appropriate post-hoc test (effectively, a measure of carry-over for some stimulus pairings).

Similarly, Davidenko *et al.* (2011) anticipated the possibility of adaptation effects yielding spurious prototype effects in a block design study similar to Loffler *et al.* (2005). The authors mitigated the effect of stimulus variation upon their measurement by matching variability within block while varying prototypicality across block. For block-design studies, this method is a appropriate mitigation of the confound. We describe below additional, and more comprehensive, responses to this confound.

Principally, covariates for prototype and neural similarity effects (more generally, direct and carry-over effects) should be included in the same general linear model. As our empirical example demonstrates, the presence of either effect may then be measured after accounting for the confounding regularities that exist in the order of stimulus presentation (see also, e.g., Kahn *et al.,* 2010). The use of a fully counter-balanced stimulus sequence (Aguirre *et al.*, 2011) is crucial, as this allows the two types of effects to be estimated efficiently and without bias (De Carlo *et al.*, 1990).[2]

A limitation of this solution is that the degree of correlation between direct and carry-over effects can become substantial, particularly in fMRI studies which are affected by the

[2] Our paper which introduced the notion of simultaneous modeling of direct and carry-over effects in neuroimaging (Aguirre 2007) erroneously states that “direct and carry-over effects are orthogonal when the order of presentation of stimuli is serially first-order balanced”. While this is true for the forms of neural response considered in that paper, the current work demonstrates that it is not a true statement generally, as confounds do exist for some forms of response.

temporal filtering properties of the hemodynamic response. Careful design of stimulus sequences can enhance power for measurement of carry-over effects (Aguirre et al., 2011), improving the ability to model carry-over effects for measurement or removal. More broadly, a model that includes both carry-over and direct effects may be assessed with an omnibus F-test without negative consequences of correlation within the covariates. This test would reveal that the neural signal does code for the stimuli or their relationship without determining the relative contribution of these effects.

One means of avoiding this issue is through the design of stimulus spaces. For example, stimuli drawn from a circle within a two-dimensional space are not subject to this confound, as every stimulus is equidistant from the (prototypical) center of the space. Of course, a downside to such a design is the inability to present a stimulus in the center of the space, thus precluding the study of norm-based coding.

Finally, experimental design may be used to minimize the presence of carry-over effects within the data. For example, a sparse fMRI design with long inter-stimulus intervals (e.g., greater than 6 seconds) would both plausibly reduce adaptation effects and reduce the degree of confound. Stimulus masking may be used to similar effect. The effectiveness of these measures for the reduction of carry-over effects would be an empirical question, with measurement of the effect subject to the same confounds discussed.

3.8 Discussion

We have explored a particular confound in the study of perceptual variation and stimulus representations. While individual neuroimaging studies have sought evidence for either prototype or adaptation effects in relation to neural encoding schemes**,** we find that these have the potential to be confused. We further demonstrate this confound empirically, and find that spurious effects can arise with incomplete modeling of fMRI data.

While this confound does not negate the existence of claimed neural effects, it may call into question their interpretation. As we have recommended, researchers interested in solely norm-based or adaptation effects have several avenues toward isolation of their effect of interest. We would argue, however, that instead of striving to solely mitigate one effect or the other, a more holistic perspective toward neural coding effects and their interaction could be useful.

As we have discussed, the concept of a prototype effect is related to a differential absolute response to a central stimulus relative to an extreme one. We can classify these prototype effects as a type of *first-order* effect of representation - one related to the unique neural response to a stimulus. Adaptation effects pertain to the overlap between distributed neural responses and arise from comparisons of stimuli - a type of *second-order* effect of representation.

The preservation of both the uniqueness of each stimulus, and the relationships of stimuli to each other is thought to be a primary goal of perceptual representation. In psychology, these two features are captured in the concept of a *stimulus space* - a representation of both stimulus identity and stimulus relationships. Stimulus spaces posit a unique location for each exemplar, with distances within space as an index of stimulus similarity. In relating this conceptual space to *neural* representations, it is possible to reframe prototype and adaptation effects. *First-order* effects of prototypicality speak to the position within a stimulus space, and *second-order* effects of similarity (e.g. adaptation) speak to distance. The concept of a stimulus space goes further than these simple features; such spaces have *topology*. In his classic psychological observation, Amos Tversky (1977) highlighted that similarity relationships were prone to asymmetry, particularly in comparisons involving prototypes. For example, most observers judge an ellipse to be more like a circle, than a circle to be like an ellipse. Such asymmetries can be understood as slopes in the surfaces of a stimulus space, biasing an otherwise equal metric relationship toward one direction. These slopes create a surface topology oriented about the prototype that has both *second-order* and *first-order* consequences.

On the second-order, surface topology would result in biases of *adaptation* effects in the direction of the prototype. Just as metric similarity has a neural adaptive effect, so does this asymmetric bias. As demonstrated by Kahn *et al.* (2010) in an ERP study, differential adaptation in the N170 and N250 evoked potentials occurs for comparisons of prototypical and extreme faces dependent upon the order of comparison - a *second-order* effect of prototype. Notably, this finding would be impossible without simultaneous modeling of *first-order* prototype effects and *second-order* similarity effects.

On the first-order, surface topology would result in differential elevation of two points in the stimulus space. As the topological slopes are oriented toward prototypical stimuli, the elevation, and therefore absolute neural response to prototypical stimuli would be reduced relative to extreme stimuli - exactly the prediction of norm-based encoding models. Thus thinking in terms of stimulus space *topology* highlights several of the dominant effects of representation currently researched.

The cohesion of the topological stimulus space model lends itself to one more avenue of investigation, namely the *dynamics* of this space, and the interactions which might drive changes in topology. Importantly, Panis *et al*. (2010) offered evidence in favor of a dynamic prototype effect. Were the dual effects of prototype and stimulus similarity modeled in parallel, it might have been possible to disentangle both the first- and second-order effects of stimuli, as well as the interaction of the two in driving the dynamics of the other.

We believe this is a promising area of investigation. It is possible that *second-order* effects of neural adaptation are instrumental in the molding of prototype effects in the short and long term. One prediction is that the pattern of neural adaptation effects across the course of an experiment changes in concert with the emergence of the dynamic prototype, as suggested by Panis *et al.* (2010).

The conclusions of this paper are therefore two-fold. In regard recent neuroimaging studies, we highlight a confound of stimulus effects which draw into question existing interpretations. We also suggest a more cohesive approach to investigating neural stimulus spaces that enables study of the dynamics of perceptual representation.

4 A Single Temporal Integration Mechanism Unites Neural Adaptation and Prototype Formation

4.1 Abstract

What information is encoded in a cortical visual representation?

That visual representations are distributed across the ventral temporal cortex is well established. fMRI adaptation studies demonstrate these neural codes are modulated by the perceptual similarity of sequential stimuli. Studies investigating prototype-based coding effects propose that neural responses are proportional to distinctiveness from a central reference, or prototype. In existing fMRI work, these two effects are considered independently. We propose here that these two effects arise as a consequence of a single mechanism of coding based upon temporal integration over recent stimulus history.

Using a carry-over fMRI design, we show significant neural adaptation and prototype-based coding effects in a face-responsive region of interest in the right fusiform gyrus when effects are modeled discretely. By considering these effects as extremes of a single *drifting norm* model, we find that visual representations tend to encode identity in terms of intermediate stimulus history. Looking beyond the region of interest, we demonstrate that the effect of temporal context varies smoothly across the cortex, with the modulatory effect of recent visual history extending further back in time in a posterior to anterior fashion along the right ventral temporal cortex.

These findings reframe two branches of the visual representation literature in terms of a unified encoding model. Importantly, this finding offers a perspective on the cortical

topology of visual identity representations. We discuss the implications of this gradient as an organizing principle of the ventral visual topology.

4.2 Introduction

The responses of neural populations are modulated both by systematic variations in stimulus properties, as well as by the recent history of stimuli.

The distinctiveness of a stimulus is a behaviorally relevant and much studied dimension of stimulus variation that impacts neural response. A graded increase in bulk neural response to stimuli that are increasingly different from a central, “prototype” stimulus is seen in both BOLD fMRI (Loffler *et al.*, 2005; Panis *et al.*, 2011; Davidenko *et al.*, 2011) and single unit electrophysiological recordings (Leopold *et al.*, 2006, De Baene *et al.*, 2007). These results are taken as evidence that neurons implement a “norm-based” code that represents stimulus exemplars with respect to a stored representation of a centrally positioned prototype (Leopold *et al.*, 2001; Rhodes & Jeffery, 2006). Quantitatively, the magnitude of norm-based neural responses is found to be proportional to the distance of the currently presented stimulus from the center of a multi-dimensional stimulus space from which the stimuli are drawn (Anderson & Wilson, 2005; Loffler *et al.*, 2005).

The effect of the history of presented stimuli upon the perception and neural representation of the currently presented stimulus is also a topic of great interest. Neural adaptation effects are the primary example of this influence of stimulus history as studied in electrophysiologic and functional neuroimaging studies (Grill-Spector & Malach, 2001). In these experiments, the repetition of an immediately preceding stimulus produces a reduced neural response to the subsequent presentation. Adaptation effects are found to parametrically vary with the similarity between stimuli (Drucker & Aguirre, 2009, Jiang *et al.*, 2006). Specifically, the adaptation of neural response is found to be proportional to the distance of the currently presented stimulus from the previously presented stimulus within a multi-dimensional stimulus space.

It is interesting to observe that both norm-based and adaptation effects are related to distance within a representational stimulus space, with the precise magnitude of the effect relative to the position of the prior stimulus, or to the center of the stimulus space. For norm-based studies, a recent question is how the stored prototype that resides at the center of the stimulus space is initially generated (Tsao & Freiwald 2006). Clearly, this is an effect of stimulus history, but one with a potential influence of the lifetime of sensory experiences. Recent findings, however, suggest that the formation of the central prototype is a more dynamic process operating on shorter time scales. In a study of novel abstract shapes (Panis *et al.*, 2011), norm-based neural effects were observed when distinctiveness was defined over the course of an experimental run.

The current literature would suggest that both neural adaptation and norm-based responses are properties of neural systems, each with some sensitivity to the history of presented stimuli, but operating at very different timescales. Could it be, however, that instead of two separate mechanisms, there is a single, intermediate representation of recent stimulus history that can account for both effects? What would be the nature of such an effect?

Consider the simplest example of the sequential presentation of three stimuli. According to a neural adaptation model, the response to the second stimulus will be larger proportional to its perceptual distance from the first, and the response to the third stimulus will be scaled by its perceptual distance to the second. If response modulation were to accumulate according to a monotonically decreasing function of stimulus history, the response to the third stimulus would be modulated by the distance to the prior stimulus and partially by the distance to the first - in effect a modulation relative to a norm which resides in between the two preceding stimuli. In the context of a longer stimulus presentation, the same rationale can be applied, leading to the prediction of a *drifting norm* in perceptual space. The drifting norm represents the temporally integrated stimulus history, and serves as a reference point from which the degree of neural response modulation is proportional. Indeed, any exponentially decaying model of neural adaptation is, in principle, a model of a drifting norm.

Here, we test the hypothesis that a single mechanism of temporal integration of stimulus history can account for both neural adaptation and prototype effects in fMRI measures of face perception. To test this hypothesis, we began by replicating prior findings of norm-based and neural adaptive effects in isolated models within a face-responsive region of interest. We find that both effects are present in our data. We then examine whether a drifting norm model better accounts for the variation in fMRI responses to faces than other models of stimulus history effects. We show that, within a region where both neural adaptation and norm-based effects can be measured, a drifting norm model corresponding to an intermediate temporal integration window best explains the variation of BOLD fMRI response to a stream of face stimuli. We then go on to investigate how the horizon of the observed temporal integration effect varies across the ventral cortical visual pathway.

4.3 Methods

4.3.1 Participants

A total of 41 subjects contributed data to either Dataset #1 or Dataset #2. From a total of 20 subjects enrolled in Dataset #1, one subject was discarded because of lost behavioral data, one subject for not responding in more than 15% trials, and three discarded for excessive head motion, leaving fifteen subjects (twelve right handed, ten female, aged 19-25 years) whose data were analyzed. From a total of 21 subjects enrolled in Dataset #2, two subjects were discarded for excessive head motion, leaving nineteen subjects whose data were analyzed. All subjects provided informed consent and the study protocol was approved by the Institutional Review Board of the University of Pennsylvania.

4.3.2 Scanning

Magnetic resonance images were obtained at 3.0-T on a Siemens Trio equipped with an 8-channel head coil at the Hospital of the University of Pennsylvania. T1-weighted structural images (160 axial slices, voxel size = 0.98 x 0.98 x 1.00 mm) were collected using a 3D magnetization-prepared rapid gradient-echo pulse sequence. During experimental runs, blood-oxygenation level dependent (BOLD) functional images were collected using an echo-planar pulse sequence (time repetition [TR] = 3 sec, time echo

[TE] = 3 ms, voxel size = 3.00 mm isotropic). Functional data were acquired with 64 x 64 in-plane resolution across 45 axial slices. Pre-processing of the functional data involved sinc-interpolation in time to correct for the slice acquisition order and motion corrected using least squares minimization. The MPRAGE image from each subject was reconstructed in surface space and mapped to the fsaverage template using FreeSurfer; functional data were transformed to the surface space and smoothed with a 10mm FWHM kernel.

4.3.3 Stimuli and Experimental Design

Two sets of synthetic facial stimuli were created for Datasets #1 and #2 using GenHead (version 1.2, Genemation). Each stimulus set was created with 3 primary axes with 3 points along each axis, resulting in 27 distinct stimuli. For Dataset #1, the three axes were gender, race, and internal facial features (Figure 4.1A) and for Dataset #2, the axes were skin tone, facial thickness, and facial identity (Figure 4.2C inset); stimuli in the second dataset were physically less distinct than the first dataset. All stimuli were created in an "older" and "younger" version; this orthogonal fourth dimension was used in the scanning cover task.

Stimuli were presented using a carry-over design[20]. Each trial lasted 1500 msec (1400 msec stimulus, 100 msec blank screen). Blank trials, in which no stimulus was presented, were also counterbalanced and doubled in length to 3000 msec. Stimuli, subtending 5˚x5˚ of visual angle, were back-projected onto a screen and viewed by subjects via a head coil-mounted mirror. On each stimulus trial, the presented face was randomly set to appear in the “older” or “younger” version. Subjects were directed to judge the age of each face and respond. A linear four button response box was held with both hands; a dual-thumb outer button press corresponded to a younger face judgment, and a dual-thumb inner button press to older faces. All but 4 subjects in both datasets performed above chance on this perceptually demanding attention task. Subjects were instructed to withhold a response on blank trials. The percentage of trials for which the subject failed to respond was taken as an index of poor attention to the stimuli. Prior to scanning, subjects performed a brief pilot session of twenty-five trials during which they practiced the age judgment task with the same stimuli used in the scanning experiments and received feedback.

4.3.4 Stimulus Sequence

Trial order was determined by Type-1, Index-1 (k = 28), first-order counterbalanced sequences (optimized as described in Appendix B of Aguirre 2007 [20]). For Dataset #1, the same sequence was used for all subjects, while for Dataset #2 a different sequence was used for each subject. As blank trials were doubled in duration, the total number of effective trials analyzed was 1624 (812 TRs).

The total sequence was spread over six functional scans of 141 TRs each. To allow for appropriate adaptive context to be achieved at the beginning of each run, the last 10 stimuli (5 TRs) from the preceding run were presented at the beginning of each scan (for the first run, the final 10 stimuli in the sequence were presented). The first 5 TRs of each scan were discarded during pre-processing. For the final scan, the sequence completed prior to completion, and thus the last 4 TRs of the scan were discarded. Thus, 136 images from the first five scans and 132 from the last run were analyzed.

4.3.5 Behavioral Assessment of Stimulus Similarity

Each subject performed a set of explicit judgments of similarity for the pairs of faces following MRI scanning. Each pairing of faces was rated on a scale from 1 to 10 (1 being identical, 10 being completely different). For Dataset #1, similarity ratings provided by different subjects were strongly concordant (the average correlation of each subject to the remainder of the group was 0.75). Consistent ratings were obtained for both old and young face sets (average between set correlation was 0.95). These similarity ratings were z-transformed within subject, averaged across subjects, scaled between 0 and 1, and entered into a multi-dimensional scaling analysis in Matlab (Mathworks, Natick, MA). The first three dimensions of the result defined the behavioral stimulus space (Figure 4.1A) used for subsequent analysis. This 3-dimensional MDS solution explained 70% of the variance in the behavioral data. A second behavioral experiment using implicit measures of perceptual similarity (inverse reaction time in a discrimination task) yielded a highly correlated result (correlation of implicit and explicit group similarity matrices = 0.79).

The same analyses were conducted for face stimuli for Dataset #2, though only 14 of the scanned subjects were available to provide similarity data. For the explicit similarity data, the average correlation of each subject to the average of the group was 0.67. A second behavioral experiment using implicit measures of perceptual similarity (inverse reaction time in a discrimination task) yielded a highly correlated result (correlation of implicit and explicit group similarity matrices = 0.86).

4.3.6 ROI Definition

A one-sample group (across subject) mean GLM was run on the main effect of stimuli versus the blank screen. The result was masked with the FreeSurfer fusiform label, and then the top 800 vertices in the right hemisphere selected to define a region of interest (ROI). The average signal across vertices was obtained for each subject and then further examined.

4.3.7 General Linear Model

The central hypothesis of the study is that the neural response to the sequence of face stimuli is best explained by a drifting norm model. In this conception, neural response is modulated in a carry-over fashion by the similarity of the presented stimulus to a single drifting norm - the current integrated representational context of the neural system. The drifting norm ($\boldsymbol{x}$) is expressed as a position within the 3-dimensional MDS stimulus space defined by behavioral measures. For the first trial, the drifting norm is arbitrarily set to the center of the stimulus space. The norm position is then updated by the presentation of stimuli ($\boldsymbol{s}$), which draw the norm towards the position of the current stimulus in the perceptual space. The degree to which the norm drifts in response to a stimulus presentation is determined a scaled decay rate (μ):

$$\boldsymbol{x}_{\mathrm{n}} = \mu \boldsymbol{s}_{\mathrm{n}} + (1-\mu)\boldsymbol{x}_{\mathrm{n-1}}$$

where $\boldsymbol{x}$ and $\boldsymbol{s}$ are coordinates in the three-dimensional MDS perceptual space, and μ is between zero and unity. The sequential application of this equation to a particular ordering of stimuli with known perceptual similarity yields a covariate that predicts the continual modulation of neural response to the presentation of the stimuli, dependent upon the scaled decay rate (μ) selected. When the μ is set to zero (Figure 4.1C, lower left panel), the resulting covariate is precisely a global norm model, in which the response to each stimulus is modeled as proportional to the distance of the stimulus from the center of the MDS perceptual stimulus space; there is no predicted effect of sequential stimulus transitions upon neural response. When μ is set to one (Figure 4.1C, lower right panel), the covariate produced is a 1-back adaptation model, in which the predicted response is proportional to the distance of the prior stimulus to the current one in the MDS-reconstructed stimulus space; the system has no memory of stimuli past the last stimulus presented. Intermediate values of μ correspond to varying degrees of temporal integration across the sequence of stimuli.

We modeled the data from the region of interest for each subject using a set of models, each of which incorporated a drifting norm covariate with a scaled decay rate (μ) ranging from 0 to 1 in steps of 0.05. Each covariate was mean centered and scaled to have unit variance. Additional covariates (common to all models) fit the main effect of the stimulus presentation versus the blank trials and the effect of stimuli presented following a blank trial. All neural model covariates were convolved with a canonical hemodynamic response function (Aguirre et al., 1998). Additional regressors were included to account for global signal, between-scan variation, and subject-specific spikes.

4.3.8 Whole Brain Mapping

For each value of μ, we combined subjects individual surface maps in FreeSurfer through a group analysis, resulting in 21 group surface maps. We then combined these into a single surface map containing at each vertex the μ with the largest beta from the 21 possible. From this surface map, we cropped all vertices where the largest corresponding beta or the main effect of faces were negative. We also removed any vertices where the p-value associated with the largest beta was less than 0.5 (Figure 4.2A). To quantify the anterior-posterior trend present on the whole brain map, 15 circular ROIs with a 15-vertex radius were plotted on the fsaverage surface in a continuous line, running anteriorly along the ventral temporal lobe from the temporal pole (Figure 4.2A, top panel overlay). On the group maps of Figure 4.2A, the average μ value within each of the 15 ROIs was evaluated for each of the 21 drifting norm models, in order to reveal the trend from posterior to anterior visual areas (Figure 4.2B). The standard error of the mean at each anatomical position was estimated by resampling (bootstrapping). The set of subjects was sampled with replacement up to the total number of available subjects. The mean value from this resampling was retained and 1000 such means were recorded across bootstraps. The standard deviation of this set of means provides an estimate of the standard error of the mean of the measure.

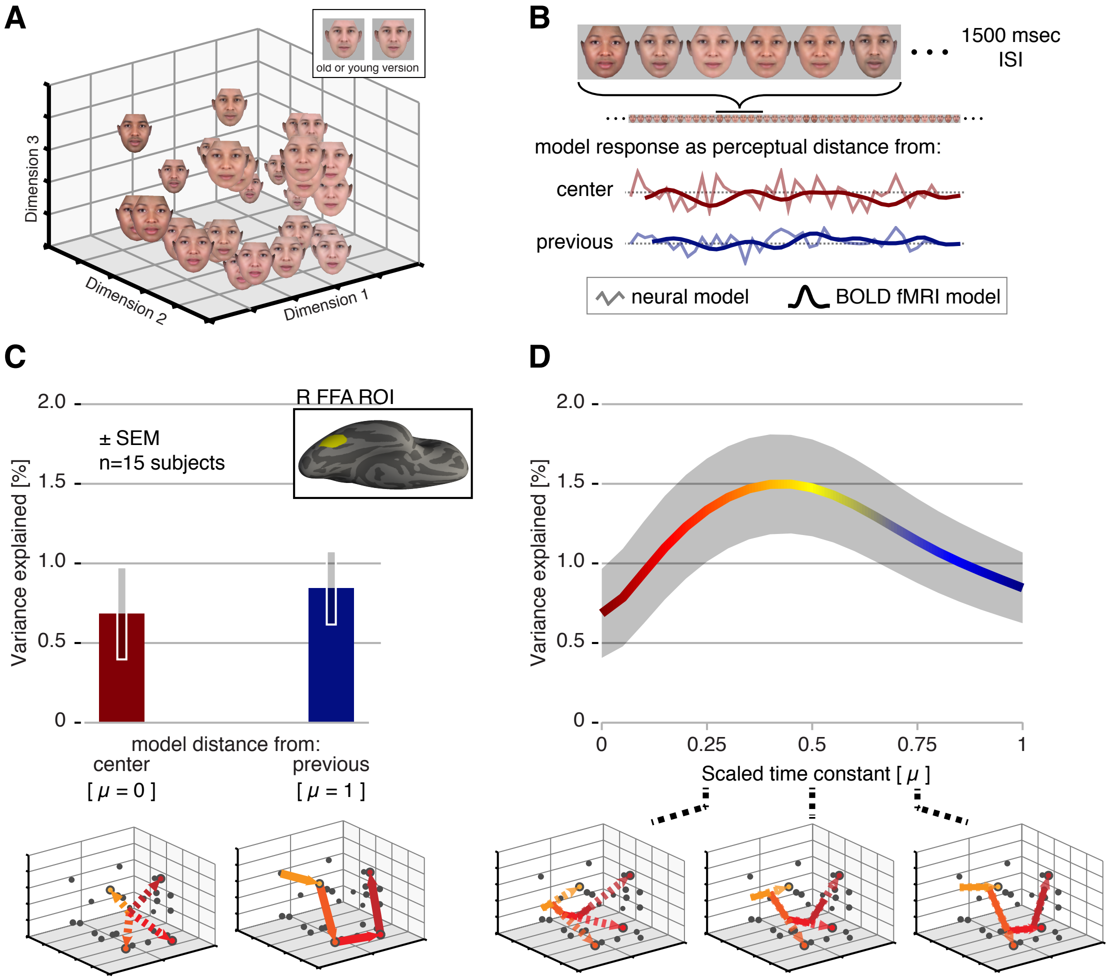


Figure 4.1: Stimuli, design, and region-of-interest analysis
(a) Faces varied in identity, race, and gender. Each was generated in an “older” or “younger” version (inset). Separate ratings of pair-wise face similarity were used to generate the perceptual similarity space. (b) During scanning, subjects observed a continuous stream of face stimuli and indicated for each if it was the “older” or “younger” version. Neural responses were modeled using continuous covariates corresponding to the perceptual distance of the current stimulus from either the previous face or the center of the stimulus space. (c) Average across-subject regression coefficients (n=15) within the ROI for norm-based and adaptation effects within an across-subject, face-responsive region of interest (ROI) in the right fusiform gyrus (inset). Stimulus space figures below illustrate how neural response to each face is modeled with respect to the center of the perceptual space or to the prior stimulus for five example stimulus transitions. (d) Fit to the neural data within the ROI for a range of decay rates (μ). The stimulus space figures illustrate how the drifting norm (solid vectors) is increasingly responsive to the most recently presented stimulus at ever higher decay rates.

4.4 Results

We obtained BOLD fMRI data while subjects viewed a continuous stream of face stimuli (Figure 4.1B). Separately, the perceptual similarity of the set of 27 faces was measured for each subject. These perceptual judgments were found to be very similar across subjects, and thus combined to produce an average perceptual similarity space (Figure 4.1A).

We identified across subjects a face-responsive region of the right fusiform gyrus (Figure 4.1C inset). We focus on the right hemisphere in this study given evidence that sensitivity to face variation is substantially stronger in the right as compared to the left hemisphere (Mur et al., 2012). This region has a consistently large, main effect of neural response across all the faces presented during the experiment. We then sought to replicate prior findings of "norm-based" and "adaptation" effects in the neural responses to faces within this region.

4.4.1 Norm-Based and Adaptation Effects

Prior studies of neural responses to faces have found that more distinctive faces evoke a larger neural response (Loffler *et al.*, 2005; Panis *et al.*, 2011; Davidenko *et al.*, 2011). We tested for such an effect in our data using a covariate in which the response to a face is modeled as being proportional to the distance of that face from the center of the perceptual stimulus space (Figure 4.1B, red curve; also Figure 4.1C, lower left panel). When modeled in this way, we find a significant norm-based response across subjects (Figure 4.1C, left bar).

Separate studies have observed that the response to a stimulus may be modulated by the similarity of the immediately preceding stimulus (Drucker & Aguirre, 2009, Jiang *et al.*, 2006). We tested for such an effect in our data using a covariate which models the neural response to a stimulus as proportional to the distance between the current face and the immediately preceding face (Figure 4.1B, blue curve; also Figure 4.1C, lower right panel). When modeled in this fashion, we find a significant proportional adaptation response across subjects (Figure 4.1C, right bar).

An initial interpretation of these results is that adaptation and norm-based effects coexist in the responses of neural populations. While we have argued elsewhere that these effects may be mistaken for one another in measurement (Kahn & Aguirre, 2012), we consider here a deeper connection: that the observed norm-based and adaptation effects are simply different measurements of a single temporal integration mechanism operating over recent stimulus history.

4.4.2 A Single Temporal Integration Mechanism

We next tested the hypothesis that stimulus history is temporally integrated into a single reference point in the stimulus space, which we term the “drifting norm”. The presentation of a new stimulus draws the drifting norm towards the new stimulus. The degree to which the current stimulus changes the drifting norm can be described by a scaled decay rate (μ). At the boundaries, a fully “elastic” neural system updates the drifting norm to match the last presented stimulus (μ=1), behaving exactly like a 1-back neural adaptation effect (Drucker & Aguirre, 2009, Jiang *et al.*, 2006), while a “rigid” neural system retains the norm at its initial position regardless of presented stimuli (μ=0), behaving like a norm-based effect (Loffler *et al.*, 2005; Davidenko *et al.*, 2011). This model is isometric with an exponential integrator over stimulus history with varying decay rates, here simplified to scaled values of μ between 0 and 1.

We tested this idea by estimating the best fitting parameter of temporal integration (μ) for each subject within the fusiform region of interest. A set of models, with scaled decay rates ranging between 0.0 (rigid norm) and 1.0 (fully elastic 1-back adaptation) were assessed. For each coefficient, we determined how much variance in the neural data was explained by the model, and obtained the average and variability of this measure across subjects.

The amount of variance explained in the neural data across subjects as a function of μ within the right fusiform is shown in Figure 4.1D. As the drifting norm is allowed to relax from rigid to fully elastic, the fit to the neural data steadily improves and then declines, reaching a peak at an intermediate value of μ (0.40). This finding suggests that face-responsive neurons maintain an integrated representation over the last several stimuli to which the currently presented face is compared.

While this result is consistent with a single temporal integration mechanism, it alone cannot adjudicate between a single integration mechanism over intermediate temporal history and a neural system that truly incorporates both adaptation and a norm-based effects. We addressed this question using a cross-validation procedure in which the data from *n-1* subjects were used to estimate the scaled decay rate and the individual contributions of norm-based and adaptation effects. For the *n*th subject, we then compared the variance explained by a model combining the norm-based and adaptation covariates with the estimated weights, with the variance explained by the drift covariate. In all 15 subjects, the proportion of variance explained by the single mechanism model was greater than the variance explained by the combined model (R^2 single mechanism model = 0.0070±0.0016, R^2 combined model = 0.0042±0.0011, both mean ± SEM).

4.4.3 Drifting Norms Across the Visual Pathway

The analyses to this point confirms the validity of the drifting norm model within the right fusiform gyrus, a region known to be responsive to face stimuli and where findings of norm-based coding are usually described. Visual stimuli, however, evoke broad neural responses across ventral visual cortices. While particular category-selective visual areas have peak responses to preferred stimuli, the identity and relative similarity of visual objects may be decoded from the broader responses (Kriegeskorte *et al.*, 2008). We therefore examined how neural response modulation based on temporal context varied across ventral visual cortex. Broadly, we anticipated that earlier visual areas would show greater modulation on the short time-scale and ever more anterior visual areas would integrate over longer time-scales (Hasson *et al.*, 2008).

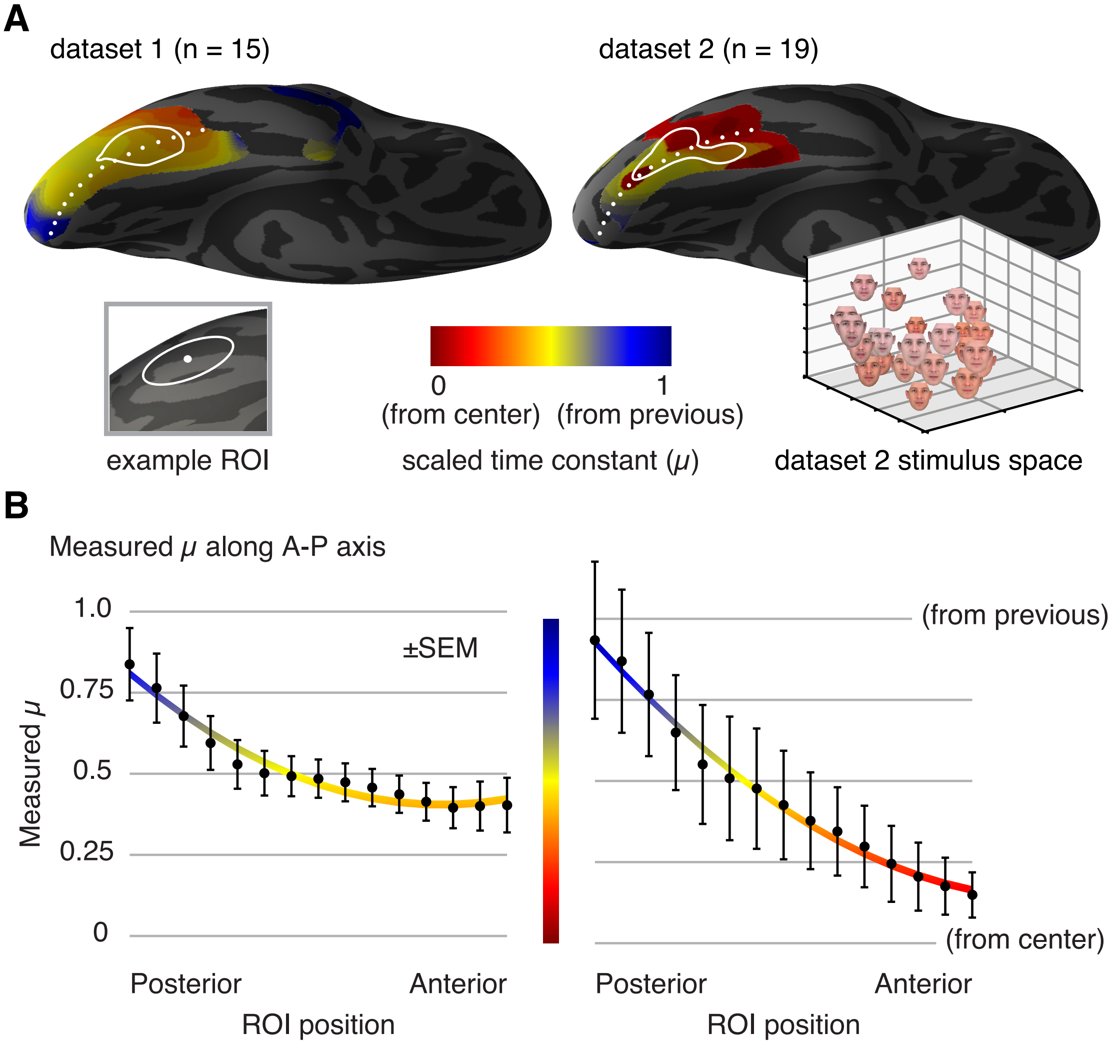


Figure 4.2: A gradient of temporal integration
(a) Optimal scaled decay rate (μ) across subjects at each ventral occipito-temporal cortical site for dataset 1. White points indicate the centers of the ROIs sampled in panel b. The stimulus space for this dataset is inset. (b) Plot of across-subject average, regional μ from the posterior to anterior ventral occipito-temporal cortex, fit with a second-order polynomial. Averages taken from cortical regions of the size and positions indicated in panel A. (c) Optimal μ across subjects as measured in a separate dataset. The stimulus space for this dataset is inset. (d) Plot of across-subject, average, regional μ for the second dataset.

We measured the μ that corresponded to the best-fitting drifting norm model separately for every vertex across the right cortical surface. We examined only those vertices that showed a positive main effect of face presentation compared to a blank screen across subjects and a positive effect of stimulus dissimilarity. We found that near the occipital pole, in the vicinity of the calcarine sulcus, the measured μ was 1.0, indicating that the neural response within this earliest visual area is modulated only by the immediately preceding stimulus. Moving anteriorly along the ventral cortex, the measured μ declines, indicating an ever greater degree of integration of stimulus history (Figure 4.2A).

We quantified this gradient by defining a path running from the occipital pole anteriorly along the ventral surface of the occipito-temporal cortex. Figure 4.2B plots the average peak μ within patches of cortex centered at each of 15 points along this path, revealing the increasing integration of temporal context in ever more anterior visual areas. A second degree polynomial fit confirmed the ever longer degree of integration of stimulus history, between early visual areas and higher level areas.

Figure 4.2 demonstrates a posterior-to-anterior gradient of temporal effects along the ventral visual stream, in which increasingly long integrated windows of stimulus history modulate the response to a presented stimuli. The gradient (Figure 4.2A) is described by the peak decay rate (μ) which best explains the neural signal measured at each patch of cortex. This map dispenses with the relative strength of these effects, however, and additional perspectives are necessary.

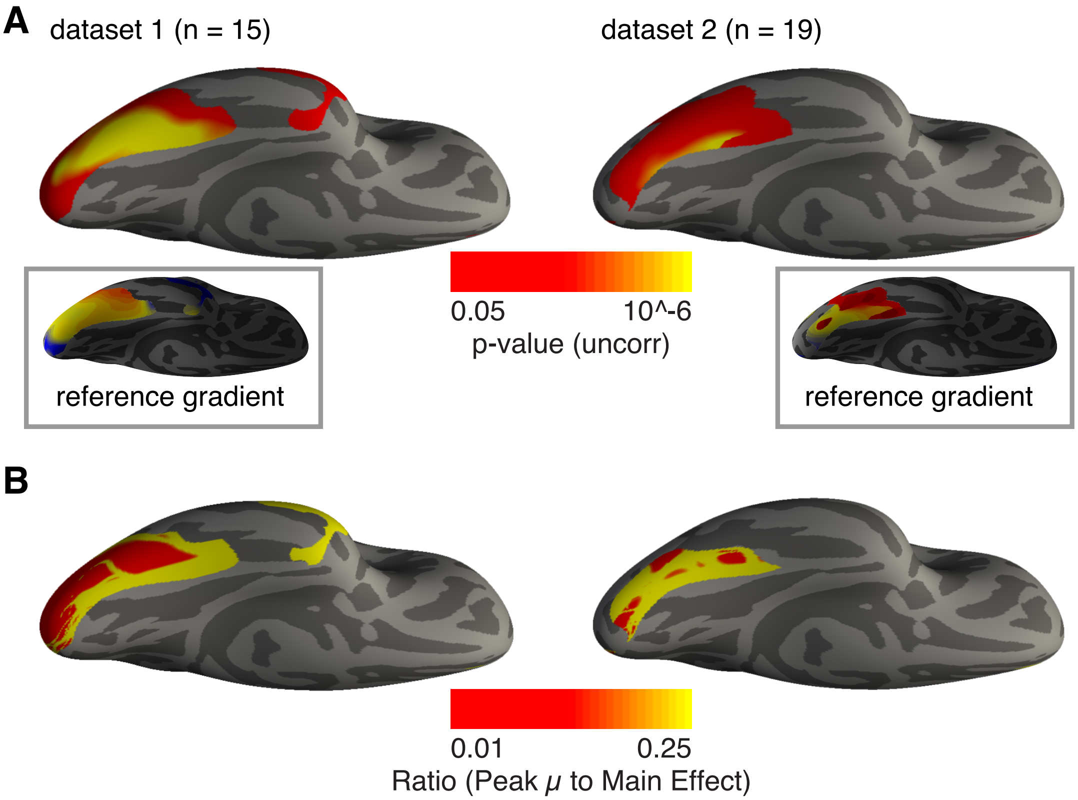


Figure 4.3: Strength of measured temporal integration models on the cortical surface (a) The p-values associated with each measured decay rate model at all vertices measured. (b) The ratio of the variance of the measured decay rate model and the variance of the main effect model at each vertex. These are cropped to vertices with values greater than 0.01 and fully saturated for values above 0.25. Values >> 1 are cropped. Since the covariates for the main effect and all temporal modulatory models are scaled to one, this is equivalent to a ratio of the beta values for each.

A surface map of the p-values (Figure 4.3A) for the peak drift model (Figure 4.3A, inset) illustrates the significance of the effects at each vertex. It appears the medial edges near the fusiform portion of the gradient show stronger effects. Complementarily informative is a plot of the ratio of the variance of the modeled peak temporal effect against the variance of the modeled main effect (Figure 4.3B). For regions where the main effect of facial stimuli is large, this ratio will be relatively reduced (red). This is true overlapping the FFA region of interest, selected specifically for having large main effects of stimuli. For regions where both the p-value of the peak temporal model is relatively low, and the temporal modulation covariate is larger relative to the main effect of stimuli, it is increasingly possible that the temporal modulatory effect model is capturing noise. For more anterior regions, this seems to be the case.

4.5 Discussion

In this study, we investigated modulatory mechanisms of the neural response to facial stimuli - specifically the relationship between norm-based representation and short-term neural adaptation. We proposed a single mechanism based upon temporal integration to unite these two effects, where the neural response to a stimulus is modulated as a function of distance to a "drifting" norm. In turn, the norm's position is updated by the recent visual experience at a rate determined by a scaled decay rate (μ), which can be directly related to an underlying exponential temporal integration function.

For μ close to zero, the drifting norm model behaves like a short-term neural adaptation effect, modulating neural responses as a function of the shortest-term stimulus history. For μ close to one, the drifting norm model behaves like a norm-based effect. A neural population that has an intermediate timescale of temporal integration, when probed for either adaptation or global norm effects in a neuroscience experiment, will be found to have both, resulting from the correlation of both adaptation and global norm predictors with the true single effect. We confirmed that an intermediate drifting norm model can better describe the data than either the norm-based or 1-back adaptation model, and used a model comparison and cross-validation approach to prove that the drifting norm model outperforms a combination of both earlier models. Lastly, we extend our findings beyond our initial region of interest to investigate the broader topology of temporal contextual encoding. We observe a gradient of increasing temporal integration from primary visual cortex extending anteriorly along the ventral surface of the right temporal lobe. This result suggests a hierarchy of temporal integration at ever higher levels of extrastriate cortex.

Merely at a conceptual stage, dynamic encoding based on recent temporal context solves a major problem of putative norm-based representation models: how is the brain supposed to “know” the arbitrary center of an unseen set? Within a unified framework of a drifting norm, any non-zero μ will position the drifting norm near the center of the space within a reasonable number of stimulus exposures, regardless of the initial set-point of the neural system.

Neural adaptation is an ubiquitous phenomenon in sensory systems, occurring at many - if not all - levels of the sensory hierarchy, and is thought to be mediated by various mechanisms beyond merely the short-term adaptation effects here described. Similarly, the realm of norm-based coding models - many of which postulate a special status for a “global” norm - reaches beyond the narrow form of dynamic norm-based effects (Davidenko *et al.*, 2011, Panis *et al.,* 2011) approached in this study. The literature on all these phenomena is extensive, yet still evolving. Our results cannot settle many questions

- for instance that of global norm-based encoding - yet these findings offer a novel inclusive perspective on these effects as well as exciting avenues for future investigation. For example, as we have previously recognized, short-term neural adaptation and norm-based coding effects as previously studied can be confounded in measurement (Kahn & Aguirre, 2012). Our work here resolves this confound, suggesting both may be manifestations of a single modulatory effect. With respect to postulated differential responses to global norms, with careful experimental design, these could be tested while controlling for the modulatory effect of recent stimulus context. Further, it could be that the modulatory effect here demonstrated might represent a mechanism by which the representational geometry of visual cortex is lastingly modified by visual experience. This possibility could be tested by carefully steering the path of the drifting norm to repeatedly cross different non-central locations in a perceptual space for different subjects - while maintaining mean stimulus exposure - and then probing the distributed response after hours or days. Lastly, it may be that multiple drifting norms could be maintained by the visual system, coding for different stimulus features or categories, with distinct windows of temporal integration distributed across the visual hierarchy.

Neural encoding based on temporal context is conceptually linked to the work of Hasson et al. (2008). In an elegant study using short video stimuli presented intact, in reverse, or temporally scrambled, the researchers demonstrate a hierarchy of “temporal receptive windows” across the cortex, with ever more anterior regions sensitive to ever longer snippets of video content. Our finding of a monotonic posterior-anterior gradient of temporal integration is consistent with this result. Hasson et al. (2008) show consistent differences between regions in the duration of temporal integration, suggesting different portions of cortex may have inherent temporal windows over which visual information is encoded.

The finding of differential modulatory effects of stimulus history for increasingly anterior portions of the right temporal lobe is interesting. Previously, the large majority of studies showing fMRI adaptation effect for faces have focused on the fusiform face area or immediately adjacent (e.g. Andrews & Ewbank, 2004; James & Gauthier, 2006; Fang *et al*. 2007; Harris & Aguirre, 2010). There exist studies demonstrating differential fMRI adaptation effects (with a temporal component) in anterior regions, though for non-face stimuli (Epstein *et al*. 2008). The closest analog in the face literature to a different posterior to anterior adaptive difference is demonstrated by Weiner *et al*. (2010), though the regional difference demonstrated was between posterior- and middle-fusiform (pFus and mFus, respectively).

This relative lack of precedent could draw into question the reliability of our measurements for increasingly anterior regions. Face identity information has been observed to be resolvable using distributed pattern analysis in the right anterior temporal lobe (Kriegeskorte *et al*, 2007), even in cases when a fusiform ROI is unable to distinguish them. Taking into account the possibility that distributed patterns & neural adaptation index different features of neural information (Drucker & Aguirre, 2009; Epstein & Morgan, 2011), it could be the case that the coarseness of identify information increases from posterior-to-anterior, in accordance with Kriegeskorte *et al*. (2007). If this were the case, we might expect that the apparent underperformance of our modulatory models (Figure 4.3) anterior to the FFA resulted from the relative weakness of individual vertex modeling on a "coarser" neural coding region. Two aspects of our data support this interpretation. First, the series of ROI analyses (Figure 4.2B) demonstrate a smooth gradient of temporal integration. These ROIs will aggregate over small effects to give a stable estimate of regional behavior. Phrased differently: though the reliability of individual vertex results may decline for increasingly anterior regions, the central tendency within region of interest should be stable. Second: we observe a difference in the anterior pattern of the temporal gradient between datasets. Given that the second dataset was generated using a more tightly clustered stimulus set, we might expect that it would be less susceptible to the effects of coarseness in increasingly anterior regions. Both of these observations are post-hoc; future work is necessary to understand the relationship of our measures to the underlying neural code. Regardless, the possibility that both temporal history effects and relative neural code coarseness vary along similar trajectories could be deeply informative.

These findings could account for the variable effect of stimulus history in neuroscience and behavioral experiments. A single modulatory mechanism based on temporal integration offers a tidy explanation for the different findings, in strikingly similar studies, of either norm-based or adaptation effects (Panis *et al*., 2011; Kahn & Aguirre, 2012). Different paradigms, response conditions, and subject instructions arguably probe the cortical gradient of temporal integration at different levels, producing behavioral effects ranging from short-term perceptual adaptation to the dynamic formation of prototype representations.

Our study is also not suitable for answering the latent question of whether the determining factor in temporal integration of stimulus history is time or number of trials. Indeed, our findings from with a region-of-interest show that our measured modulatory effect has a half-life of either 3-4 stimuli or 4-5 seconds. In single-unit studies in early sensory systems (e.g., the fly H1 visual neuron or mouse retinal ganglion cell; Fairhall *et al.*, 2001; Wark *et al.*, 2009), the timescale of integration was found to vary with the timescale of stimulus changes. This suggests that neural systems can scale temporal integration to the context of stimulus variability. Interestingly, a cascade of decaying exponential integrators has this property (Wark *et al.* , 2007), which could correspond to the gradient of temporal integration we observe along the ventral temporal cortex.

In neural and behavioral terms, the implications of a moving norm are open to interpretation. One on hand, a moving norm can be thought of as implementing gain control, preventing large transitions in the perceptual space from saturating neural responses by constantly readjusting its reference point. Alternatively, the norm in our model can be understood as a ”prediction” of the upcoming stimulus given previous sensory evidence, with the evoked signal corresponding to an error signal. This possibility is in agreement with the predictive coding hypothesis (Friston & Kiebel, 2009), and could be tested empirically via novel experimental design. Regardless of the underlying theory, we believe that our model is a useful and didactic description of the cumulative effects of neural response modulation.

Major advances have been made in recent years toward the goal of mapping the representational geometry of visual cortex. Distributed neural population responses, as measured by BOLD fMRI, have been shown to contain information regarding the large-scale perceptual similarity relationships of broad categories of visual stimuli (Kriegeskorte *et al*. 2008), and recently it was suggested that this representational geometry is relatively uniform and ubiquitous across ventral visual cortex (Cohen & Alvarez, 2014). Our results presented here, focusing on the nuanced dynamics of distributed representation, speak to the questions of why visual information should be so broadly distributed across the ventral stream, and how processing unfolds over a seemingly uniform scaffold.

A representational neural state is a product of both the underlying geometry and modulatory effects, such as those of temporal context demonstrated here and likely others, such as task demands. Moving forward, it may be productive to consider visual neural encoding in terms of a manifold, a dimensioned encoding structure with temporal and task-based response characteristics. In this framework, the currently studied representational geometry of visual cortex (Kriegeskorte *et al*., 2008)) would be a primary map projection of the manifold, and higher dimensions could scaffold the dynamic properties of perception. The interactions of stimulus properties, task demands, or cortical areas with the representational geometry could be collected as parallel charts of the manifold through careful experimental manipulations. Investigations such as these would allow for a more nuanced understanding of what information is encoded in a given neural state, and thus facilitate the future pursuit of neural decoding.

5 An ERP Index of Visual “Sameness” is Altered in Autism Spectrum Disorder

5.1 Abstract

In the search for a neural etiology of autism spectrum disorder, Leo Kanner’s original observation (1943) of an “insistence on sameness” in the autistic phenotype provides a promising avenue of inquiry, prompting? the question: “how does the brain evaluate sameness?” This question is far from trivial, though one answer lies in how neural systems relate current sensory information to recent stimulus history. Advances in neuroimaging have lent perspective on this evaluation, suggesting that neural responses to visual stimuli are modulated by similarity to a prior that reflects temporally integrated stimulus history.

In this study, we ask whether this modulatory effect is altered in individuals with autism spectrum disorder. As the timescale of temporal integration can be related to the moment-to-moment changes in the visual environment as well as the generalization of stimulus relationships (i.e., across a longer timescale), we hypothesized that autism would be marked by response modulations based on increasingly immediate stimulus history. Using electroencephalography, we measured event-related potentials to morphed face stimuli in 54 young adults (27 with ASD). By modeling the timescale of the evoked response modulations, we find that ASD is marked by effects of information accumulated over shorter temporal windows. We relate this difference in neural processing to recent theories of perceptual functioning in ASD.

5.2 Introduction

A broad collection of neurocognitive theories of ASD have implicated perceptual processing as a locus of dysfunction. Early and enduring theories, such as the weak central coherence account (Frith 1989, Happé & Frith, 2006), the enhanced discrimination and reduced generalization hypothesis (Plaisted, 2001), and the enhanced perceptual functioning account (Mottron *et al*., 2006), highlight perceptual features of ASD as driving imbalances in cognitive functions and abilities. These proposed differences commonly take the form of a bias toward local processing at the expense of global understanding, difficulties with abstracting or generalizing commonalities (the cognitive style of 'missing the forest for the trees'), and a more veridical perspective on the world. Though the predictions of these theories have been met with mixed success experimentally, they have succeeded in keeping the realm of perception central to discussions of autistic etiology.

As theoretic accounts of ASD have evolved, the discussion of perception has become grounded in theories of sensory systems - including the process of neural representation. Neural representation is the means by which sensory information is encoded in patterns of neural activity, and the nature of a neural representation determines the information available to a neural system for guiding learning and behavior. Though the idea that differences in neural representation could drive ASD symptomology is not new (Gustafsson, 1997; McClelland, 2000), recent theoretical accounts of ASD have re-emphasized the potential explanatory importance of sensory encoding / decoding disruptions in ASD. These recent neurocognitive accounts approach the sensory process holistically, considering it, for instance, in terms of a Bayesian inference calculator (Pellicano & Burr, 2012), or a predictive coding system (Lawson *et al*., 2013; Van de Cruys *et al*., 2014).

While theoretic approaches possess intellectual heft, they require neural correlates to establish an experimental foothold in the brain. The work presented here begins with a model of neural response modulations that are related to stimulus representation. By beginning with a neural model, we remain agnostic to the broader theoretic frameworks, but benefit from the proximity of our model to the brain.

The central measure of this study is an extension a long-studied feature of sensory systems: neural adaptation - an attenuation of neural response to repetitions of similar or identical stimuli (Grill-Spector & Malach, 2001). Recently, it was demonstrated that neural adaptation effects extend in time - that is, the modulatory effect on evoked responses depends upon stimulus history beyond merely the most recent stimulus seen (Mattar *et al*., in preparation). For continuous, serial presentations of stimuli, sensory history is best modeled as being temporally integrated into a *drifting norm*. The drifting norm is in essence a continually updated “prior” used as a reference for neural response modulation.

The character of the drifting norm depends upon the window of temporal integration. For long temporal integration windows, the drifting norm behaves like a prototype - residing near the center of a representational stimulus space. For increasingly short temporal windows, the drifting norm reflects moment-to-moment stimulus history. At the shortest extreme of the model, the drifting norm is updated by each new stimulus, with the resulting modulation behaving effectively like a 1-back neural adaptive effect. Using fMRI, Mattar *et al*. (in preparation) demonstrate BOLD signal modulations in the fusiform gyrus of neurotypical participants are best fit by an intermediate window of temporal integration.

We use the same framework (Mattar *et al.*, in preparation) in this study to model modulations to the evoked responses to faces in electroencephalography (EEG). The temporal character of the “drifting norm” is captured in a single parameter - the scaled decay rate (μ). Moment-to-moment updating of the drifting norm (akin to 1-back adaptation effects) is based on a maximal decay rate ($\mu = 1$) while generalization of central tendency (akin to a prototype-referenced modulation) corresponds to a null decay rate ($\mu = 0$).

Our prediction for ASD in this case is straightforward; we anticipated a shift toward modulation of neural responses based on *more recent* stimulus history. The reasons for this are several-fold. For one, the temporal integration model is related to the establishment of a prototype - a representation of the central tendency of a stimulus space. From a certain perspective, this is the essence of weak central coherence (Frith, 1989; Happé & Frith, 2006) or the reduced generalization hypothesis - the ability to abstract a sense of central tendency. Behavioral studies of ASD support our expectation. A reduced ability to form perceptual prototypes has been suggested in ASD (Klinger &

Dawson, 2001), and face adaptive after-effects have been shown to be reduced in ASD (Pellicano *et al*., 2007). Given that these effects both build up over time and rely on a sense of central tendency (Rhodes & Jeffery, 2006), a shift toward neural modulation based on a more recent window of stimulus history could be mechanistic to these behavioral observations in ASD.

We measured event related potentials (ERPs) while participants viewed a continuous stream of morphed faces while performing an orthogonal oddball task. We examined first the modulations of the P200 component of the evoked response, given that it is known to be sensitive to effects of recent stimulus history (Kahn *et al*., 2010). Modeling the extremes of the drifting norm framework ($\mu = 0$ and $\mu = 1$) demonstrates a group difference in this component, which we build upon by examining the full range of decay rates. We then investigate modulatory effects across the duration of the measured evoked response in a component-independent manner. The group difference we demonstrate suggests that the shift toward moment-to-moment modulations we observe in ASD has unexpected cascading effects on the time course of face processing. We go on to discuss the implications of increasingly moment-to-moment modulation of neural responses in ASD on recent neurocognitive models of the disorder.

5.3 Methods

5.3.1 Participants

Fifty-four young adult subjects (27 with ASD, 27 neurotypical - NT) were included in this study.

Participants were recruited from the Philadelphia area and surrounding region. All participants were compensated for their time, travel costs, and a bonus (up to $10) for task attention. All subjects earned the entire bonus. Informed consent was obtained from all participants after a complete description of procedures, and the study design was reviewed and approved by the Institutional Review Board (IRB) of the Children's Hospital of Philadelphia.

The two groups were matched on FSIQ, handedness, and gender ratio. The groups differed significantly by age (a mean difference of 1.43 years that was not deemed clinically relevant). See Table 1 for a subject information.

Volunteers and parents of volunteers were screened over the phone to rule out the presence of Axis I disorders, uncorrected auditory or visual impairments, or significant medical or neurological abnormalities or injuries. Participants with ASD had their diagnosis confirmed by current symptom presentation (Autism Diagnostic Observation Schedule[ADOS], Module 4; Lord *et al*., 2009) and autism-specific developmental history (Autism Diagnostic Interview [ADI-R]; Lord *et al*., 1994). Clinical assessments were administered, scored, and interpreted by clinical psychologists or doctoral level trainees under supervision of clinical psychologists. Full Scale IQ above 75, determined by Weschler Abbreviated Scale of Intelligence (WASI-II; Wechsler, 2008), was used as an inclusionary criterion.

An additional 6 participants provided EEG data but were excluded from analysis. Two subjects were excluded prior to pre-processing: one subject (ASD) was excluded based on clinician uncertainty over diagnosis, and a second subject (neurotypical) was excluded due to a setup error during data collection. Four subjects (two with ASD, two neurotypical) were excluded based on excessive artifacts during pre-processing (exceeding 40% of modeled trials).

Table 5.1: Participant characterization

	ASD (N = 27)	**TDC (N = 27)**	
	Mean(SD)	Mean (SD)	Significance
Age in years	19.64(1.61)	21.07(2.38)	t(52)=2.59, p=0.012
FSIQ	109.96(17.44)	113.19(11.29)	t(52)=-0.81, p=0.42
Gender	22 male, 5 female	21 male, 5 female	Fisher's exact test p = 1.0
Handedness	25 right, 2 left	25 right, 2 left	Fisher's exact test p = 1.0

5.3.2 Stimuli

Two female faces with neutral expressions (identities 06F and 10F) from the NimStim stimulus set (Tottenham *et al*., 2009) were used to create a linear morph continuum. The raw faces were aligned and resized to match inter-pupillary distance and location in Adobe Photoshop CS5 (Adobe, San Jose, CA). The images were desaturated, major blemishes were removed, and the faces were cropped of external features with a 3-pixel feathered oval boundary. Using MATLAB (MathWorks, Natick, MA), the mean luminance of the faces was equilibrated. Wireframe templates were fitted to the images using JPsychomorph (http://cherry.dcs.aber.ac.uk:8080/wiki/jpsychomorph) and the faces were morphed in 11 equal steps. After morphing, the images were balanced for mean luminance once again. Six faces (Figure 4.1A) from the 11 original steps were used in the experiment (steps 2, 3, 5, 7, 9, 10). In addition, four male faces to be used as targets were drawn from the NimStim set (23M 27M 28M 35M) were prepared in the same manner, but were not morphed. The stimuli from the NimStim set are not authorized for publication, all figures use representative identities created using the same methods.

5.3.3 ERP Stimulus Presentation

The experiment consisted of a total of 1082 trials. Stimulus order was set using a series of 3 first-order counterbalanced ($k = 18$, $n = 2$) de Bruijn cycles (Aguirre *et al.*, 2012). The 18-element sequence allowed for counterbalancing of the 6 morphed stimuli and 3 ISI durations. Nine break periods occurred evenly through the experiment, during which the experiment paused until the participant clicked to continue. Ten "target" trials occurred at pseudorandom points during the experiment. After each target and break, the five preceding trials were repeated to "warm-up" the carry-over effects of interest. For the beginning of the experiment, the five final trials were presented as warm up. Three separate versions of the full sequence were generated and counterbalanced across participants to reduce the potential for higher-order stimulus order effects. All target and warm-up trials were excluded from the analysis, leaving only the 972 trials from the three cycles used in the main analyses.

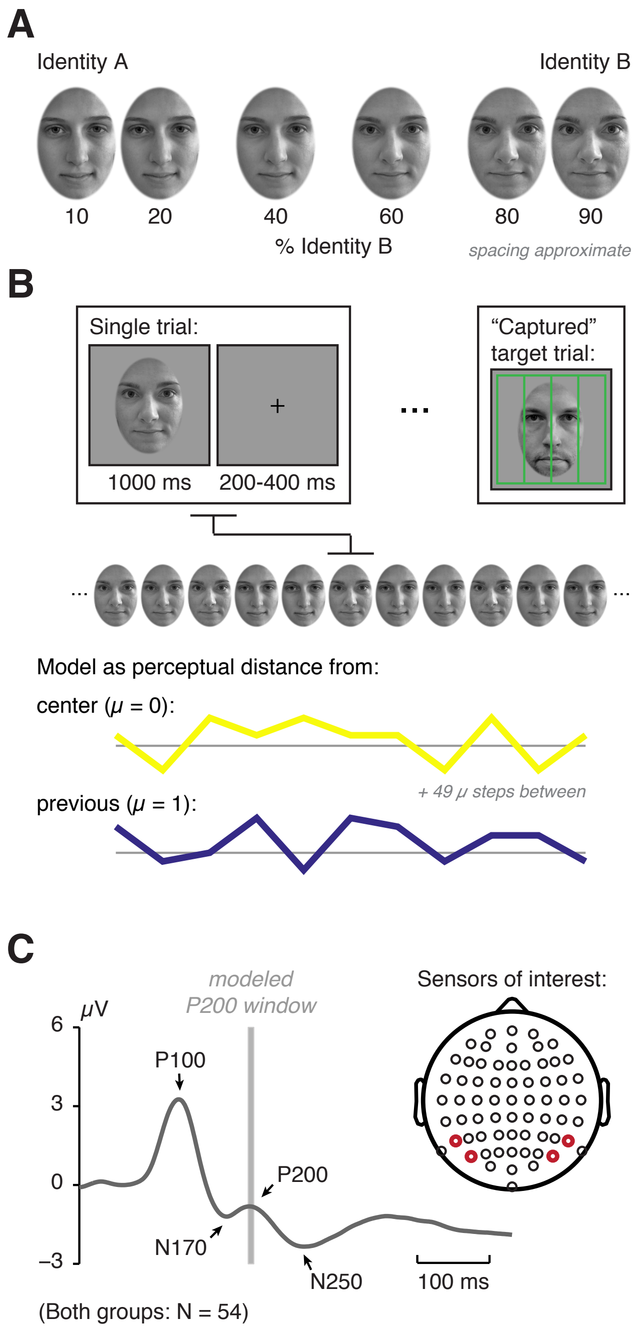


Figure 5.1: Experimental design & modeling
(a) Stimuli were six exemplars (unequally spaced) drawn from a linear morph continuum of two identities. The actual anchor identities were taken from the NimStim stimulus set (Tottenham *et al*., 2009); example stimuli shown are two of the author's friends. (a) Experimental design and modeling. Morphed stimuli were presented continuously while subjects monitored for the rare appearance of "robbers" and responded with a button press (example "robber" is representative). Trials were modeled as the perceptual distance from a reference point which varied as a function of the scaled decay rate (µ). (c) Grand average evoked response across the sensors-of-interest (SOIs). Four bilateral SOIs (inset) were selected using an independent localizer. The average waveforms across subjects shows four canonical components within the 600 ms epoch. The P200 window indicated was used for component-of-interest analysis.

Each trial (Figure 5.1B) consisted of a stimulus presentation subtending 5.7˚ x 8.6˚ of visual angle on a gray background for 1000 ms, followed by an inter-stimulus interval of 200, 300, or 400 ms during which a black fixation cross was shown (Figure 5.1B).

Stimuli were presented on a 19” ViewSonic G90fb CRT display using EPrime 2 (Psychology Software Tools, Inc.) situated at eye level 100 cm from participants. During the experiments, participants performed an orthogonal task (“catch the robbers”) in which they responded via button press to the appearance of any of four male faces (the “robbers”) among the female distractors. Participants were trained on a version of the task using different distractor faces immediately prior to the main experiment in order to familiarize themselves with the target faces. Participants were generally amused by the task. Participants responded to targets via mouse click and responses were collected in EPrime 2. For successful identification of targets, a green set of “bars” would appear over the face (Figure 5.1B), indicating a correct response. For “false alarm” responses, a red frame would appear around the face to indicate the incorrect response.

Following the main experiment, participants completed a short passive-viewing functional localizer consisting of faces, cityscapes, and objects (72 exemplars each, subtending 12.5˚ x 12.5˚ of visual angle) in a random order. In the localizer, stimuli were presented for 400 ms with a jittered ITI between 800 and 1200 ms. The localizer was used in order to independently assess “sensors of interest” for use in the main analysis (Liu, Harris, & Kanwisher, 2002).

5.3.4 ERP Data Collection

Data were collected using the BioSemi ActiveTwo system (http://www.biosemi.com/products.htm) with 64 active electrodes with sintered Ag-AgCl tips fitted into sized head caps. Additionally, two electrodes with a flat 4-mm pallet were placed on the mastoid processes bilaterally with adhesive stickers and used as references for data import. Electrical offsets were verified prior to initiation of data collection & kept below 25 μV for all channels. Continuous data were sampled at a 512 Hz sample rate using the default 1/5 sample rate low-pass filtering (http://www.biosemi.com/faq/adjust_samplerate.htm). Participants were seated alone in an adjacent room from the experimenter with the door between ajar. Verbal encouragement was kept uniform and offered during breaks (e.g. "doing great, keep it up.") in order to assess subject attentiveness.

5.3.5 ERP Pre-Processing

All EEG data were pre-processed offline using MATLAB and and the EEGLAB toolbox version 9 (Delorme & Makeig, 2004). Data were imported into EEGLAB directly, mastoid channels were indicated as references and excluded. The data were re-referenced to the average signal across the 64 cranial channels, and a 40 Hz low-pass filter was applied. The continuous data were separated into 700 msec epochs (100 ms pre-stimulus onset and 600 post) corresponding to individual stimulus presentations. Epochs were baseline corrected (100 ms pre-stimulus onset). Sensors-of-interest were identified using the independent localizer data via a point-to-point t-test comparing the face and house conditions in the latency range of the N170 and N250. Sensors-of-interest (Figure 5.1C, inset) were selected if they were identified as significant in a majority of subjects. Artifacts were identified only within the sensors-of-interest using a ±50 μV threshold across the full epoch. Trials containing artifacts within the SOIs were removed from analysis. Groups did not significantly differ on frequency of artifacts (ASD: mean 10.28%, std 9.48%; NT: mean 7.32%, std 9.65%).

For the initial analyses (Figure 5.2), the P200 component of the evoked response was defined across participants. A grand average waveform was generated for all included trials and participants (Figure 5.1C). The P200 was defined as the peak value within a search window of 200 to 300 ms post-stimulus onset. The P200 amplitude for individual trials was evaluated as the mean of the 5 data-point (9.76 ms) window about this time point (Figure 5.1C, grey column).

For the analysis investigating the entire evoked response epochs (Figure 5.3), each time point of the evoked response across epochs was modeled individually. Grand average waveforms were generated for each group by averaging all modeled epochs within subject and then across subject (Figure 5.3A). At each time point, the peak μ model was determined within group. The plot of the percent variance explained for the peak model (Figure 5.3B) is used a clipping mask for a color plot demonstrating the peak μ. For Figure 5.3C, the waveforms were exported from MATLAB as EPS files and were manipulated in Adobe Illustrator CS5 for visualization. The raw paths were set to 12 pt line widths and outlined automatically by Adobe Illustrator. The resulting outline was used as a clipping mask for a columnar color plot indicating the peak μ for each time point within group after statistical correction.

5.3.6 General Linear Modeling

All analyses relied upon modified general linear models (GLMs). For every GLM here reported, the data modeled corresponded to a single narrow time window of the evoked responses across trials. The GLM would contain a covariate modeling the effect of interest, the effect of inter-stimulus interval as a nuisance covariate, and a unit offset term. Following calculation of the ß values on each covariate, the percentage of the residual variance explained by the covariate of interest would be calculated as follows. The variance of the modeled effect of interest (the variance of the covariate of the effect of interest multiplied by its ß value) would be divided by variance of the original data after ISI effects were removed. This denominator was calculated as the variance of the difference between the original data and the modeled effect of ISI (the covariate for ISI multiplied by the beta). This proportion is multiplied and reported as percent residual variance explained. All covariates of interest (described below) were mean-centered and scaled to have unit variance. Lastly, as a convention, the sign of the covariates was reversed to correspond to direction of modulation (Kahn *et al.*, 2010).

For the modeling of adaptation-like ("from previous") effects a covariate was created modeling the dissimilarity of the prior face to the current one on each trial (Figure 5.1B, blue line). These were set to have value for each trial corresponding to a linear distance within the linear stimulus space (with the six morph faces positioned at 1, 2, 4, 6, 8, and 9).The resulting covariate was mean-centered and scaled to have unit-variance. This covariate was identical to the drifting norm covariate with $\mu = 1$.

For the modeling of norm-like "from center" effects (Figure 5.1B, yellow line), a covariate was created modeling the distance of each stimulus within the linear stimulus space from the center of the space (e.g. the difference between any given face position and the average of the face positions). The resulting covariate was mean-centered and scaled to have unit-variance. This covariate was identical to the drifting norm covariate with $\mu = 0$.

Following from Mattar *et al*. (in preparation), we recognized that adaptation and norm-based effects represent narrow windows onto a single effect model of modulation based upon temporal context. The "drifting-norm" is expressed as the reference point from which the perceptual distance of the currently presented stimulus is measured. The drifting norm position is updated on each trial as a function of the scaled decay rate (μ, between 0 and unity) using the function: $Xn = Xn-1 + \mu(Sn - Xn-1)$, where Sn is the location of the most recent stimulus and Xn the location of the drifting norm on a given trial, both as positions within the linear morph space.

For each "drifting norm" covariate, the value on any given trials was set to the distance between the given stimulus and the drifting norm on that trial. 51 covariates were generated corresponding to values of μ between 0 and 1 in steps of 0.02. Each covariate was mean-centered and set to have unit variance.

5.3.7 Peak-μ Waveform Correction

For the analysis investigating the entire epoch of the evoked response, it was necessary to correct for both the number of comparisons within group & also evaluate whether group differences were significant. This was handled with a two-fold test of significance.

Within each group, the full array of % residual variance explained values [time points x μ] was tested for significance. The results were FDR correct at $\alpha < 0.05$ for each group.

Separately, group differences were assessed by subtracting the full array of residual % variance explained values [time points x μ] for neurotypical participants from that of the ASD group. To evaluate these differences for significance, a bootstrap resampling the two groups with replacement was performed 1000 times and the difference between each bootstrap at every point in the [time point x μ] array evaluated. The standard deviation of the bootstrap differences at each point was taken as the SEM of the corresponding point of the veridical difference map and used to assess significance. The resulting p-values were FDR corrected at $\alpha < 0.05$.

For the plots of the full waveforms in each group (Figure 5.3C), the time points rendered in color demonstrated significant % residual variance explained after FDR correction within group, as well as a significant corrected group difference in % residual variance explained for the time point and μ rendered. This double-correction suggests the modulations rendered represent both significant within-group and across-group effects.

5.4 Results

In this experiment, we collected EEG data while participants with ASD and neurotypical controls monitored a continuous series of linearly morphed faces for oddball targets ("robbers"). We modeled modulations of the evoked response to faces in our data based upon a "drifting norm" (Mattar *et al.*, in preparation) framework, in which response amplitude is altered as a function of distance from a reference point reflecting temporally integrated stimulus history. The variable of interest in this framework is the timescale of this temporal integration - how freely does the norm drift? At one extreme (a scaled decay rate $\mu = 0$), stimulus history is broadly integrated & thus the norm remains fixed at the center of the stimulus space. At the other extreme ($\mu = 1$), the drifting norm is updated completely to match the most recent face. Since a longer timescale of integration corresponds to the generalization of a perceptual prototype - a process argued to be altered in ASD (Klinger & Dawson, 2001) - we hypothesized that our participants with ASD would demonstrate modulations based upon more recent stimulus history.

We first modeled effects within the P200 component of the evoked response to faces, as this component has been previously shown to exhibit modulations based upon stimulus history (Kahn *et al.*, 2010). Based upon these results, we then investigated modulations of responses in a component-agnostic manner across 600 ms of the evoked response.

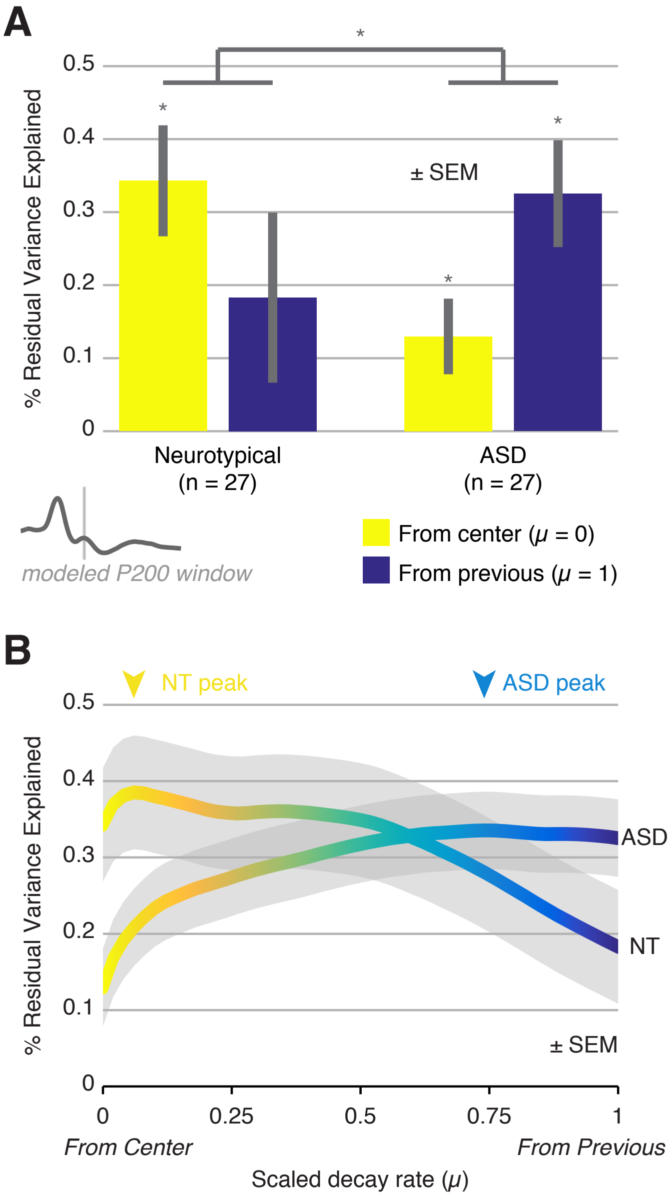


Figure 5.2: Analysis within the P200 component
(a) The amplitude of the P200 component across trials was modeled first as a function of distance from a central "norm" (yellow bars) or from the distance of the prior face to the current one (blue bars) in the linear stimulus space. A significant interaction effect of model by group was observed. Error bars correspond to the between-subject SEM. (b) The amplitude of the P200 component was next modeled using 51 "drifting norm" models corresponding to scaled decay rates (μ) between 0 and 1 in steps of 0.02. The neurotypical group (NT) demonstrated a peak μ of 0.06 (yellow arrow) while the ASD group had a peak of of 0.74 (blue arrow).

5.4.1 Analysis Within the P200 Component

The P200 component of the evoked response has been demonstrated to exhibit amplitude modulations based on recent stimulus history, and so we began by modeling the amplitude of this component. A grand average waveform (Figure 5.1C) was obtained across all participants and groups. The center of the P200 component was identified as 237.89 ms after stimulus onset, and the amplitude of the response was set as a mean of the five time points around this peak.

Following from the analyses in Mattar *et al*. (in preparation), we first targeted modulations from the extremes of the drifting norm framework, which correspond to a "1-back" neural adaptation effect - modulation as a function of current stimulus distance from the stimulus immediately preceding in time, and a "norm-based" effect - a modulation as a function of current stimulus distance from the center of the stimulus space. These effects were modeled separately in all participants, and summarized as a percentage of the residual variance explained within the P200 component after effects of inter-stimulus interval were removed. These data are presented in Figure 5.2A. In the ASD group, a significant amount of the variance was explained by the model corresponding 1-back neural adaptation ("from previous", $t(26) = 4.45$, $p < 1.43e\text{-}04$) as well as the model corresponding to a norm-based effect ("from center, $t(26) = 2.51$, $p < 0.019$). The neurotypical group showed a significant modulation of the P200 component; in contrast to the ASD group, modulation in nuerotypical corresponded to a norm based effect ("from center" $t(26) = 4.51$, $p < 1.22e\text{-}04$). A repeated-measures ANOVA comparing the groups and models demonstrated a significant interaction effect of group by model, suggesting that the boundaries of the drifting norm framework varied across diagnostic boundaries ($F(1,52) = 7.25$, $p = 0.0095$).

As the boundary conditions of the drifting norm framework appeared altered, we then investigated the full spectrum of temporal integration windows within the P200. We evaluated the percentage variance of the P200 amplitude explained by models corresponding to 51 values of the scaled decay rate (μ) across participants. These are summarized in Figure 5.2B, averaged across group. The "peak" scaled decay rate is indicated, as the max of the averaged curves for each group. The optimal scaled decay rate for neurotypical participants ($\mu = 0.06$) is smaller than that of the ASD group ($\mu = 0.74$) suggesting the P200 component is modulated in neurotypical participants based upon a *longer* window of stimulus history relative to participants with ASD. This difference can be evaluated statistically by finding the peak μ for each subject and testing the groups using a 2-sample t-test assuming unequal variance. While this between-subjects average yields a smaller difference, the group separation is significant ($t(50.25) = 2.44$, $p = 0.018$).

5.4.2 Analysis of the Full Evoked Response Time-Course

A suggested finding in ASD is the potential for phase delays in the evoked response to sensory information (Sutherland, 2010). Were the modulatory effects we are investigating to vary across time, we might anticipate that comparing groups at a fixed time after stimulus onset to demonstrate differences based upon the stage of processing and not upon the underlying window of temporal integration.

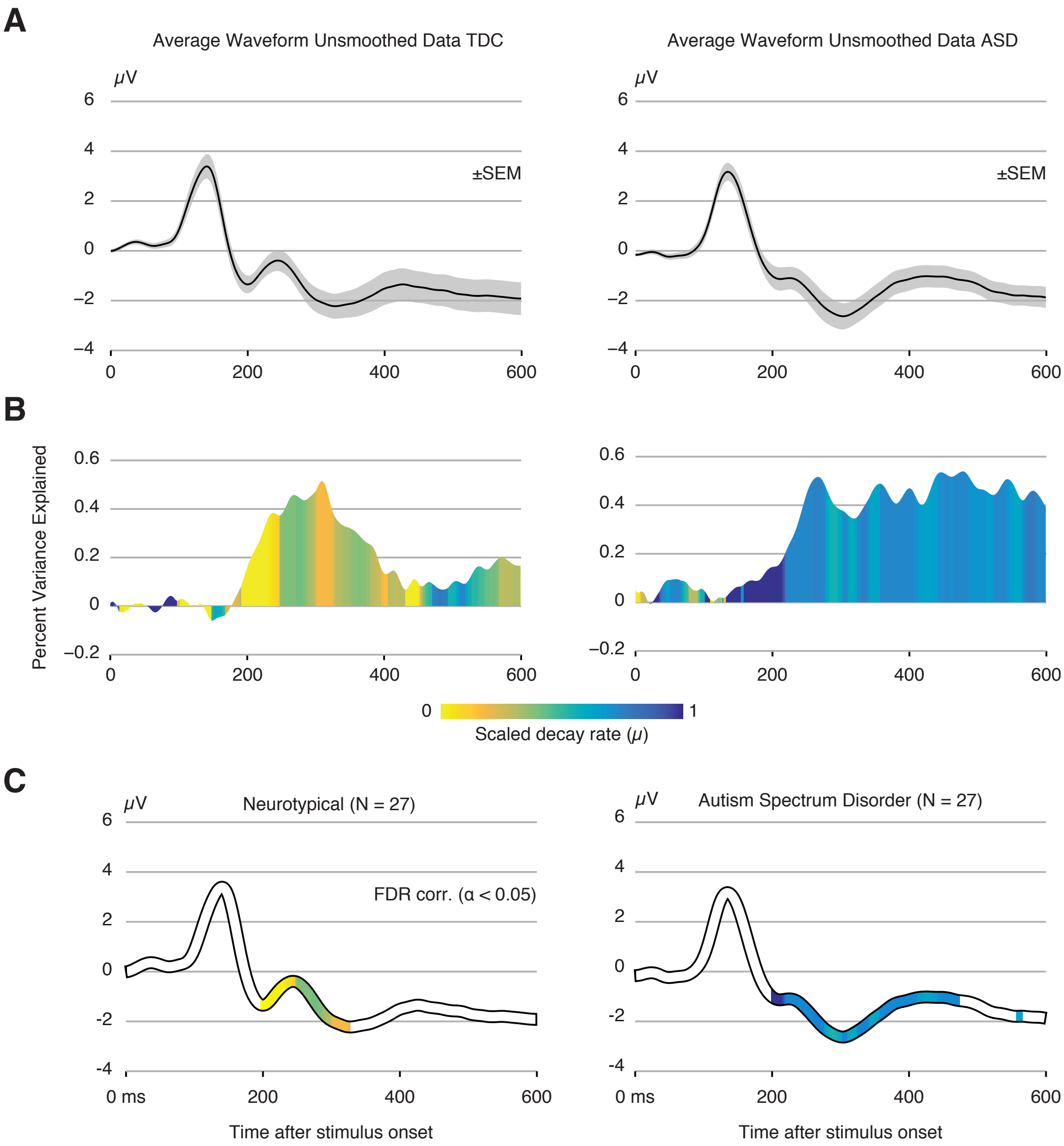


Figure 5.3: Analysis across evoked response
(a) Group average waveforms for neurotypicals (left) and ASD (right). Grey regions represent the standard error of the mean at each time point. (b) Percentage residual variance explained, after interstimulus interval effects are accounted for, labeled by the peak scaled decay rate (μ) at each time point of the 600 ms epoch. (c) Statistical analysis. Group average waveforms were labeled as in B. Time points rendered in color represented models whose % residual variance explained was significantly greater than zero (α < 0.05, FDR corrected), and for which there was a significant difference in % residual variance explained between groups for the peak μ (α < 0.05, FDR corrected).

We thus next modeled the modulatory effects of stimulus history across the entire extracted epoch (600 ms) of the evoked response. Each group waveform is plotted in Figure 5.3A. Should the modulations unfold evenly over the time course for each group but do so more slowly in ASD, it should manifest here. The percentage of the residual variance explained (after accounting for the inter-stimulus interval) by the best-fitting modulatory effect model for each group is plotted in Figure 5.3B. The color code corresponds to the peak μ of the model at that time point, evaluated as in Figure 5.2B. The statistical significance of these measurements is rendered in Figure 5.3C. The average waveform for each group is plotted with a color scale indicating the peak μ for each time point. The peak μ at each time point was tested for significance and the entire plot for each group FDR corrected for time points (308 points at 512 Hz) and drifting norm models (51 values of μ). Separately an FDR corrected group difference was calculated for each time point and drifting norm model via bootstrap resampling. Time points plotted in color represent significant within-group and between-group effects. These plots demonstrate a circumscribed modulatory effect in the neurotypical group beginning close to the center of the N170 component (196.8 ms after stimulus onset) and terminating approximately in the middle of the N250 component (325.8 ms after stimulus onset). The peak decay rate (μ) of the modulatory effect in the neurotypical group across this stage of the evoked response ranged from 0.58 to 1, with a median of 0.82.

The modulatory effects in the ASD group manifest differently across the average evoked response. The peak μ for the ASD group is consistently higher relative to the neurotypical group, ranging from 0 to 0.46, with a median of 0.26, reflecting shorter windows of temporal integration driving the modulatory effects. Further, though the onset of significant modulation begins at the N170 in the ASD group (198.8 ms after stimulus onset), the modulatory effects persist longer.

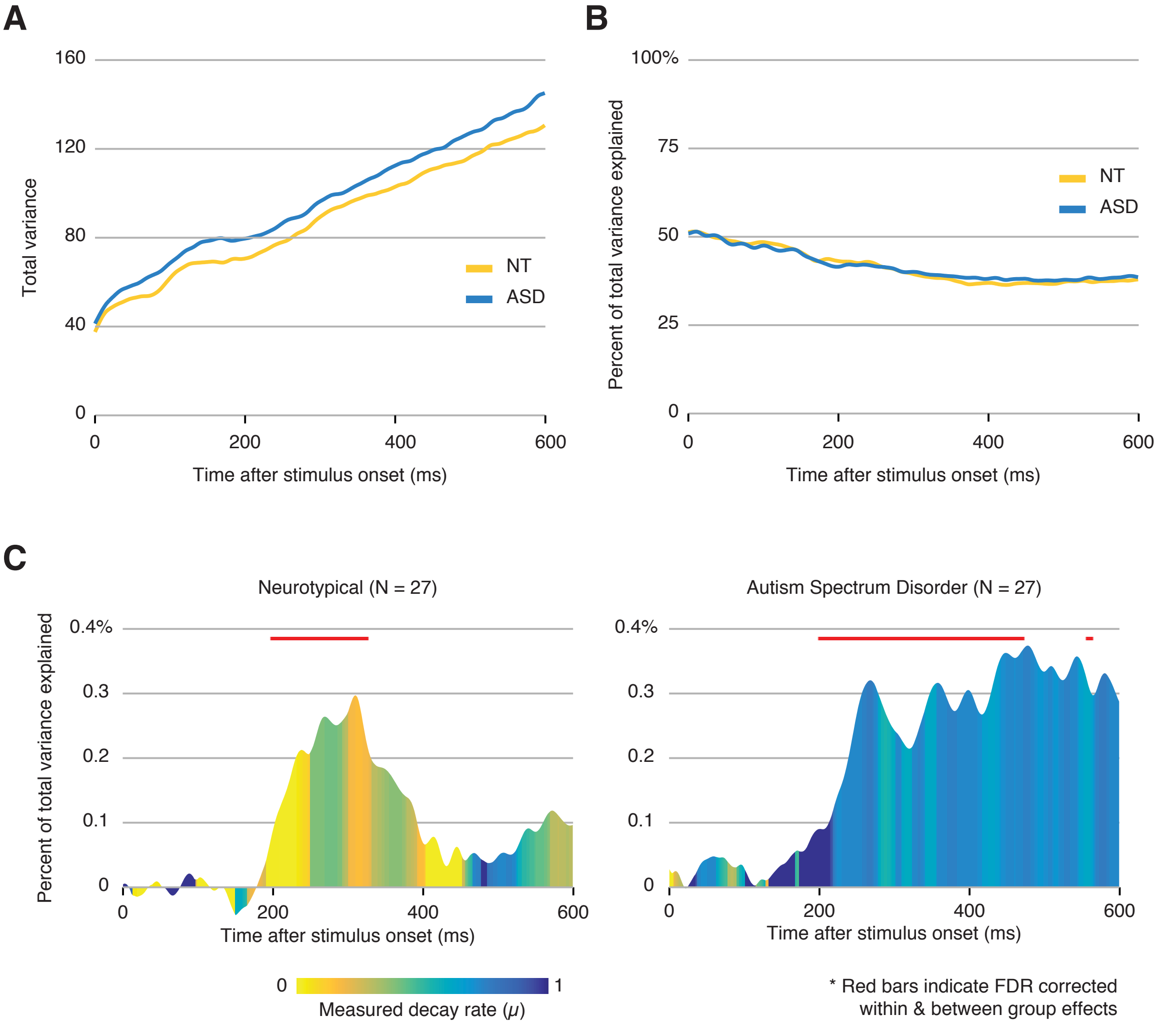


Figure 5.4: Raw variance data across the time course
(a) The average variance for each group (neurotypical in yellow, ASD in blue) for each time point of 600 msec epoch. (b) The percentage of the variance in A explained by the modeling of interstimulus interval for each group. (c) The percentage of the raw variance explained by the peak temporal history model for each group, across time.

For completeness, we also present raw variance data (Figure 5.4) including the proportion of the variance explained across time points by the effects of interstimulus interval (Figure 5.4B) for each group and the effects of the peak modulatory model (Figure 5.4C).

5.5 Discussion

In this study we investigate the time course of the evoked response to faces in autism spectrum disorder. We demonstrate that, relative to neurotypical controls, participants with ASD exhibit modulations of their evoked response based upon more recent stimulus history - in essence a more moment-to-moment neural calculation of similarity from sensory evidence to stimulus history. We demonstrate this first within the P200 component of the evoked response to faces, which has been demonstrated to show stimulus history effects (Kahn *et al*., 2010). We expand this finding by demonstrating that not only are modulations in ASD based on more recent stimulus history, but that these modulatory effects are evident for a longer period of the evoked response to faces. While we had hypothesized the former would occur in our data, the latter feature was novel. It seems very possible that a shift in the modulatory window could have cascading effects on the face processing stream.

While these results are exciting, it is important to recognize several limitations. First, we use a limited stimulus set to assess effects, so the range of potential modulations is smaller. The behavior of the drifting norm model for highly dissimilar stimuli is less well understood (Mattar *et al*., in preparation). Additionally, our findings are limited to the realm of faces. To assess whether this is a general mechanism of altered neural processing in ASD, we would want to use different stimulus modalities. Lastly, our relatively small sample size limits our ability to assess individual differences. Future work can optimize our approach to link it to the ASD phenotype.

Several recent neurocognitive models of ASD correspond well to our findings. The high, inflexible precision of prediction errors in autism account (HIPPEA, Van de Cruys *et al.*, 2014) suggests that low-level sensory information, interpreted as prediction errors, is encoded maladaptively. HIPPEA suggests that in the context of sensory responses representing deviations from expectation, ASD may be understood as an increase in precision of these prediction errors that is invariant to context. Though our model of modulatory effects on neural responses (Mattar *et al.*, in preparation) arose from an exploration of representational geometry, its conceptual basis can be readily linked to the theories of predictive encoding. The "drifting norm" here discussed can be interpreted as a rolling sensory prediction, and modulations relative to the norm (deviations) as prediction error signals. If interpreted this way, the measurements we demonstrate here corroborate HIPPEA's account (Van de Cruys *et al.*, 2014).

Related but distinct to the concept of ASD as a disorder of prediction is the proposal of a Bayesian basis for ASD (Pellicano & Burr, 2012). In this account, Pellicano & Burr suggest that to the extent sensory information processing can be modeled as a Bayesian inference, ASD could be marked by attenuated priors (or "hypopriors"), placing an overemphasis on incoming sensory information in establishing a percept. Critiques of this model (Friston *et al.*, 2013; van Boxtel & Lu, 2013; Van de Cruys *et al.*, 2013) pointed to the lack of a neural framework to explain the concept of hypopriors. Our findings could bolster this account; to the extent that the drifting norm represents a form of prior, calculation of this norm on a more moment-to-moment basis (e.g. with a higher decay rate, μ) would effectively de-emphasize generalization, and could reduce the usefulness or even establishment of a prior.

Separate from the concept of predictive coding, it was recently suggested that evoked responses in ASD are unreliable, exhibiting greater intra-individual variability of stimulus-evoked responses in fMRI (IIV, Dinstein *et al*., 2012; Haigh *et al*., 2014). It is non-trivial to assess the interactions of our findings with the account of greater intra-individual variability in ASD.

If one were to consider the "drifting norm" as an actual image prior existing in our one-dimensional stimulus space, and then to calculate the pixel-wise differences between the current stimulus and the drifting norm across the time course of the experiment, the variability of this pixel-wise difference plot over time would increase as a function of decay rate. Phrased differently, the amplitude of the signal generated by an encoding system using a given decay rate will increase as a function of decay rate. This observation reflects the fact that the visual world is prone to high frequency change inessential to an understanding material reality (we hesitate to call this noise). A neural system with a lower decay rate (μ) in essence applies a low-pass filter to this signal, dampening the features of visual change. The difference we observe in ASD, lower decay rates (μ), is a shift toward less filtering. The open question would seem to be whether greater intra-individual variability reflects a reduction in signal-to-noise or a fundamental difference in the nature of the signal encoded. We would argue in favor of the latter interpretation. The alternative view flips the relationship between IIV and μ - the argument that IIV is "upstream" of a larger decay rate for sensory evidence (μ). We think this is less likely for two reasons. If intra-individual variability is higher, it seems likely that a neural system would opt to reduce the weighting of current sensory evidence to maintain stability (lower μ for higher IIV). Also, if intrinsic noise were higher in ASD, why then is higher IIV predominately observed for stimulus-evoked signals (Dinstein *et al*., 2012; Haigh *et al*., 2014)? Generally, we feel the similarities between the increased IIV account of ASD & our findings outweigh any disagreements. We would suggest our result offers an exciting new avenue for exploring the origins of increased IIV in ASD.

In this article we have offered a perspective on the dynamics of visual evoked responses in ASD. While the modeling framework is relatively new, we feel the findings generally support a number of existing neurocognitive theories of ASD focused on perception. A great deal of research will be needed to begin to explain how the measures we assess emerge from visual networks, and how this emergence varies as a function of connectivity of the network. A mechanistic neural perspective on ASD will require more than the narrow window on the brain reported here.

6 General Discussion

The work presented in this thesis has unified two effects of neural encoding - what we called *neural adaptation* and *norm-based effects* - in terms of one modulatory mechanism based on stimulus history. The fundamental variable of this model is the decay rate (μ) which determines how much of recent sensory history is used in establishing a reference point. The updating of this reference point is done on a rolling basis - as the visual world is ever changing, so too must the drifting norm update.

Plenty of questions remain about the picture we draw of stimulus history effects on neural responses. In our experiments, we can't dissociate between the effects of recent stimulus history measured in units of time and that measured in stimulus presentations; our experiments were not designed to separate these two. One could envision an experiment where stimuli are presented for varying amount of time while also varying in metric similarity to each other. It also remains to be explored whether there is one or several references points maintained perceptually. The stimuli we used were all sampled from a single continuous space. What would happen if there were implicit groupings of stimuli? It seems likely that several local norms could be established. If so, is there an upper limit to the number of norms?

Another aspect of our experiments that gets me: how often does one really see a series of faces presented serially? While this is useful to make detailed measurements, I wonder how stimulus history effects are instantiated under natural or naturalistic viewing conditions? In this work, we begin to understand the neural calculus that is used to filter incoming visual information. How does this scale to the experience of real environment? I think this latter question is quite important when discussing the implications of our work to autism spectrum disorder. In ASD, we observe that the timescale over which the reference is calculated is shorter, corresponding to a reduction in the smoothing of sensory input. Ignoring information via smoothing isn't necessarily a net loss - there is a forest to see if you gloss over the trees. If we had a better understanding of how sensory history effects played out in full environments, we might gain an understanding of how the world "feels" to someone whose visual system smoothes less. A thought experiment I considered recently: if we could use neural measurements of sensory history effects in ASD drawn from the experience of a movie (for instance) could we apply a filter to the movie (exaggerating variation for instance) to impose similar visual cortical modulation in a neurotypical viewer? This proposal relies on a reverse inference: matching the cortical modulation doesn't mean the percept is the same. However, it might be possible to validate the method by designing stimuli that are salient to a neurotypical individual when filtered & testing whether the unfiltered stimulus is salient to someone with ASD. This prospect has potentially fun applications: if we could validate a model of how the world looks to someone with ASD, would be possible, for instance, to design a classroom which minimizes overstimulation for students on the spectrum?

If we return to the central question framed by the introduction to this thesis - how the visual system balances encoding variation and yet generalizes - we're left with a clear set questions: how is this trade-off balancing act evaluated? Phrased differently: what determines the timescale of smoothing across variation? Why is this calculation different in autism? I will speculate around a few avenues to walk in search of answers, and then walk down one a much further than I think any current research justifies.

The first avenue toward a possible answer focuses on unit characteristics. That is, on the properties and activity of neurons or perhaps cortical columns - the building blocks of the systems and streams discussed here. This avenue holds a lot of promise, as it will be most easily integrated with understanding from other fields such as genetics and animal models. The second focuses on network qualities of sensory systems - how does the distributed system appear to behave.

This second avenue has been of particular interest to me through the development of this work. One way to conceptualize visual representation is a distributed mapping of a perceptual space onto a neural network. The modulatory activity related to sensory history we discuss here could represent topological alterations to this representational surface. The gradient of modulatory activity we observe across the cortex could be understood as different dynamics of this topology - short windows of temporal integration could be like a drum which bounces back immediately after deformation, whereas longer windows are like memory foam that takes longer to reshape.

It should be noted that this finding of different temporal characteristics across the cortex has been echoed in other work using different methods. Uri Hasson used the term “temporal receptive windows” to describe the amount of time different patches of visual cortex seems to care about (Hasson *et al*. 2008). Recent work in monkeys suggested *intrinsic timescales* existed in hierarchies across cortex (Murray *et al.* 2014). The interaction of sensory information and time seems like it could be a fundamental organizing principle of cortical activity. If we were to extend the gradient we observed in Chapter 3, we might glimpse the possibility (as in Murray *et al.* 2014) of a brain-wide map of timescales. Of course, we also observe that this gradient seems to shift a bit based on stimulus properties - perhaps our map could bring in the different possible temporal-integration states of each patch of cortex into a chart across stimulus conditions. What else might cause it to vary? Let's add more maps.

What we quickly realize in this search for maps is that the measures needn't be limited to the modulatory timescales we've spent this whole thesis exploring. Countless other neural calculations unfold across cortex. Receptive fields vary in size & location. Sensory modalities as well. The early findings of fMRI demonstrated functional areas such as the fusiform face area (FFA), a patch of cortex which prefers faces to other classes of visual objects dot the cortex as well.

By analogy these are state-by-state maps in geography; perhaps our timescale gradient is something akin to mean temperature or population density. Still other maps exist. A huge wing of computational neuroscience has focused on connectivity analyses - using different imaging modalities to evaluate the strength of connections between two regions of the brain. The layering of such maps under different conditions has also been approached - Danielle Bassett and Marcelo Mattar (a collaborator of mine on the temporal integration work) at Penn have been observing differences in functional connectivity across brain areas under different task conditions.

Let's speculate what it would be like to have even more measures than receptive fields and temporal windows and connectivity profiles. What other maps might we draw? By analogy, the humanities are way ahead of us. A friend of mine during my years at Penn, Andy Fenelon, did his graduate work in demography - a field which excels at mapping. One could understand much about the United States by layering maps of population density and median income and mean social network size, for instance. Could we start to ask new questions with enough information?

A fun question I've been pondering is "why is the fusiform face area where it is?" - if we had enough maps would it be the obvious place for it to fall? (By analogy to geography, a related question is "why is the film industry in Hollywood?" - a little knowledge of climate and film history answers this).

Eventually these sort of questions and the collection of maps would demand a level of rigor. What I think we might need are new terms to describe the approach to information processing at the population level - not necessarily systems neuroscience - but truly a *crowd* or *mob* understanding within and across systems. I would suggest something like neurodemography or socioneurology. Sociology being the study of information transfer across a population, and demography the mapping of distributions of traits across a population based on factors such as geography.

In many ways, this is what the field of neuroscience is doing already - careful characterization of the brain at many levels. However, it seems to me that the useful thing (soon) would be to bring together this information in whatever and every way possible. What would follow would be the brain equivalent of what companies like Facebook or Amazon are doing currently with their users - developing cohesive, exhaustive profiles of what each user (patch of cortex) is doing across a variety of measures. This *deep profiling* is the raw material that might allow for mechanistic understandings to coalesce. And just as Amazon uses algorithms to target consumers, so too might deep learning algorithms mine neurodemographic profiles to, for instance, diagnose psychiatric disorders in a distributed way via deep learning algorithms. I think this is where the network road leads - deep profiling followed by deep learning to abstract a model of the brain's processing space.

As out there as I feel like we've reached in this discussion, I'm convinced there's much further to go. I think simply of how much I've grown over the course of my graduate school career & I'm reminded that the brain is not a static processor - the maps I could have collected (were I able) of my own brain at the outset of graduate school might not even tell me much about it today. Though I do really aspire to make sense of all this, I find that fact oddly comforting.

Bibliography

Aguirre, G.K., Zarahn, E., D'Esposito, M., (1998). The variability of human, BOLD hemodynamic responses. *NeuroImage*, *8*, 360–369.

Aguirre, G. K. (2007). Continuous carry-over designs for fMRI. *Neuroimage, 35,* 1480-1494.

Aguirre, G.K., Mattar, M.G., & Magis-Weinberg, L. (2011). de Bruijn cycles for neural decoding. *NeuroImage, 56,* 1293-1300.

American Psychiatric Association (2013). Diagnostic and Statistical Manual of Mental Disorders (DSM-V). American Psychiatric Publishing.
Anderson N.D., & Wilson H.R. (2005). The nature of synthetic face adaptation. Vision Res., 45, 1815-28.

Andrews, T.J., & Ewbank, M.P. (2004). Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *NeuroImage, 23,* 905-13.

Barlow, H. B. Possible principles underlying the transformation of sensory messages. in Rosenbluth, W. A. (ed)., Sensory Communication, (pp. 217–234). MIT Press, Cambridge, 1961.

Barlow, H.B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, *1*, 371-394.

Bentin, S., Allison, T., Puce, A., & Perez, E.. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience, 8*, 551-565.

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

Chevallier C, Kohls G, Troiani V, Brodkin ES, Schultz RT. (2012). The social motivation theory of autism. *Trends Cogn Sci., 16,* 231-239.

Davidenko, N., Remus, D.A., Grill-Spector, K. (2011). Face likeness and image variability drive responses in human face-selective ventral regions. *Hum Brain Mapp,* EPub: doi:10.1002/hbm.21367.

De Baene, W., Premereur, E., Vogels, R. (2007). Properties of shape tuning of macaque inferior temporal neurons examined using rapid serial visual presentation. *J Neurophysiol, 97,* 2900-2917.

De Baene, W., Vogels, R. (2010). Effects of adaptation on the stimulus selectivity of macaque inferior temporal spiking activity and local field potentials. *Cereb Cortex, 20*, 2145-2165.

DeCarlo, L.T., & Cross, D.V. (1990). Sequential Effects in Magnitude Scaling: Models and Theory. *Journal of Experimental Psychology General, 119*, 375-396.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods, 134*, 9-21.

Di Russo, F., Martínez, A., Sereno, M.I., Pitzalis, S., & Hillyard, S.A. (2001). The cortical sources of the early components of the visual evoked potential. *Human Brain Mapping, 15*, 95-111.

Dinstein, I., Heeger, D.J., Lorenzi, L., Minshew, N.J., Malach, R., Behrmann, M. (2012). Unreliable evoked responses in autism. *Neuron, 75,* 981-91.

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

Drucker, D.M., Kerr, W.T., & Aguirre, G.K. (2009). Distinguishing conjoint and independent neural tuning for stimulus features with fMRI adaptation. *Journal of Neurophysiology, 101*, 3310-3324.

Edelman, S. (1998). Representation is representation of similarities. *The Behavioral and Brain Sciences, 21*, 449–67.

Epstein, R.A., Parker, W.E., & Feiler, A.M. (2008). Two kinds of FMRI repetition suppression? Evidence for dissociable neural mechanisms. *J Neurophysiol, 99,* 2877-86.

Epstein RA, Morgan LK. (2012). Neural responses to visual scenes reveals inconsistencies between fMRI adaptation and multivoxel pattern analysis. *Neuropsychologia, 50,* 530-543.

Fairhall A.L., Lewen G.D., Bialek W. & de Ruyter van Steveninck R.R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature 412*, 787-792.

Fang, F., Murray, S., & He, S. (2007). Duration-dependent FMRI adaptation and distributed viewer-centered face representation in human visual cortex. *Cereb Cortex, 17,* 1402-11.

Freeman, J. B., Rule, N. O., Adams, R. B., & Ambady, N. (2010). The neural basis of categorical face perception: graded representations of face gender in fusiform and orbitofrontal cortices *Cereb Cortex, 20,* 1314–1322.

Friston, K., and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. Lond. B Biol. Sci. 364,* 1211–1221.

Friston, K.J., Lawson, R., & Frith, C.D. (2013). On hyperpriors and hypopriors: comment on Pellicano and Burr. *Trends Cogn Sci., 17,* 1.

Frith, U. (1989). Autism: explaining the enigma. Blackwell.

Furl, N., van Rijsbergen, N. J., Treves, A., Friston, K. J., & Dolan, R. J. (2007). Experience-dependent coding of facial expression in superior temporal sulcus. *Proceedings of the National Academy of Sciences, U.S.A., 104*, 13485-13489.
Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *J Exp Psychol Gen, 123*, 178-200.

Grill-Spector, K., & Malach, R. (2001). fMR-adaptation: A tool for studying the functional properties of human cortical neurons. *Acta Psychologica, 107,* 293-321.

Gustafsson, L. (1997). Inadequate cortical feature maps: a neural circuit theory of autism. *Biol Psychiatry, 42*, 1138-1147.

Haigh, S.M., Heeger, D.J., Dinstein, I., Minshew, N., & Behrmann, M. (2014). Cortical Variability in the Sensory-Evoked Response in Autism. *J Autism Dev Disord.,* Epub ahead of print.

Happé, F., & Frith, U. (2006). The weak coherence accoung: detail-focused cognitive style in autism spectrum disorders. *J Autism Dev Disord., 36,* 5-25.

Harris, A., & Nakayama, K. (2007). Rapid face-selective adaptation of an early extrastriate component in MEG. *Cerebral Cortex, 17*, 63-70.

Harris, A., & Nakayama, K. (2008). Rapid adaptation of the M170 response: Importance of face parts. *Cerebral Cortex, 18*, 467-476.

Harris, A., Aguirre, G.K. (2008). The representation of parts and wholes in face-selective cortex. *J Cogn Neurosci, 20*, 863-878.

Harris, A., & Aguirre, G.K. (2010). Neural tuning for face wholes and parts in human fusiform gyrus revealed by FMRI adaptation. *J Neurophysiol., 104,* 336-45.

Hasson U., Yang E., Vallines I., Heeger D.J., & Rubin N. (2008). A hierarchy of temporal receptive windows in human cortex. *J Neurosci., 28,* 2539-2550.

Henson, R. N. A., & Rugg, M. D. (2003). Neural response suppression, haemodynamic repetition effects, and behavioural priming. *Neuropsychologia, 41*, 263-270.

Holman, E. W. (1979). Monotonic models for asymmetric proximities. *Journal of Mathematical Psychology, 20*, 1-15.

Itier, R. J., & Taylor, M. J. (2002). Inversion and contrast polarity reversal affect both encoding and recognition processes of unfamiliar faces: A repetition study using ERPs. *Neuroimage, 15*, 353-372.

Itier, R. J., & Taylor, M. J. (2004). N170 or N1? Spatiotemporal differences between object and face processing using ERPs. *Cerebral Cortex, 14*, 132-142.

Jacques, C., & Rossion, B. (2006). The speed of individual face categorization. *Psychological Science, 17*, 485-492.
James, T.W., & Gauthier, I. (2006). Repetition-induced changes in BOLD response reflect accumulation of neural activity. *Hum Brain Mapp, 27,* 37-46.

Jiang, X., Rosen, E., Zeffiro, T., Vanmeter, J., Blanz, V., & Riesenhuber, M. (2006). Evaluation of a shape-based model of human face discrimination using fMRI and behavioral techniques. *Neuron, 50*, 159-172.

Johannesson, M. (2000). Modelling asymmetric similarity with prominence. *The British Journal of Mathematical and Statistical Psychology, 53*, 121-139.

Kahn D.A., Harris A.M., Wolk D.A., & Aguirre G.K. (2010). Temporally distinct neural coding of perceptual similarity and prototype bias. *J Vis., 10,* 12.

Kanner, L. (1943). Autistic disturbances of affective contact. *The Nervous Child, 2,* 217–250.

Klinger, L. G., & Dawson, G. (2001). Prototype formation in autism. *Dev Psychopathol, 13,* 111-124.

Kloth, N., Schweinberger, S., & Kovács, G. (2009). Neural correlates of generic versus gender-specific face adaptation. *Journal of Cognitive Neuroscience, EPub*, doi:10.1162/jocn.2009.21329

Kotsoni, E., Csibra, G., Mareschal, D., & Johnson, M.H. (2007). Electrophysiological correlates of common-onset visual masking. *Neuropsychologia, 45*, 2285-2293.

Kovács, G., Zimme,r M., Bankó, E., Harza, I., Antal, A., & Vidnyánszky, Z. (2006). Electrophysiological correlates of visual adaptation to faces and body parts in humans. *Cerebral Cortex, 16*, 742-753.

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proc Natl Acad Sci USA, 104,* 20600-5.

Kriegeskorte, N., Mur, M., Bandettini, P. (2008). Representational similarity analysis - connecting branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2,* 4.

Krumhansl, C.L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review, 85*, 445-463.

Kruskal, J.B., & Wish, M. (1978). *Multidimensional Scaling Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: Sage.

Latinus, M., Crabbe, F., & Belin, P. (2011). Learning-Induced Changes in the Cerebral Processing of Voice Identity. *Cereb Cortex*, EPub: doi:10.1093/cercor/bhr077

Lawson, R.P., Rees, G., & Friston, K.J. (2014). An aberrant precision account of autism. *Front Hum Neurosci., 8,* 302.

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

Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature, 442,* 572–575.

Liu, J., Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: an MEG study. *Nat Neurosci., 5*, 910-916.

Liu, J., Higuchi, M., Marantz, A., & Kanwisher, N. (2000). The selectivity of the occipitotemporal M170 for faces. *NeuroReport, 11*, 337-341.

Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nat Neurosci, 8*, 1386–1390.

Lord, C., Rutter, M., Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord.*, *24,* 659-685.

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., & Gotham, K. (2008). ADOS: Autism Diagnostic Observation Schedule. Western Psychological Services.

Mattar, M.G., Kahn, D.A., Thompson-Schill, S.L., & Aguirre, G.K. (in preparation). A single temporal integration mechanism unites neural adaptation and prototype formation.

McClelland, J. L. (2000). The basis of hyperspecificity in autism: a preliminary suggestion based on properties of neural nets. *J Autism Dev Disord., 30*, 497-502.

Mottron, L., Dawson, M., Soulières, I., Hubert, B., & Burack, J. (2006). Enhanced perceptual functioning in autism: an update, and eight principles of autistic perception. *J Autism Dev Disord.*, *36*, 27-43.

Mur M., Ruff D.A., Bodurka J., De Weerd P., Bandettini P.A., & Kriegeskorte N. (2012). Categorical, yet graded--single-image activation profiles of human category-selective cortical regions. *J Neurosci., 32,* 8649-62.

Murray JD, Bernacchia A, Freedman DJ, Romo R4 Wallis JD, Cai X, Padoa-Schioppa C, Pasternak T, Seo H, Lee D, Wang XJ. (2014). A hierarchy of intrinsic timescales across primate cortex. *Nat Neurosci., 17,* 1661-1663.

Nosofsky, R. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology, 23*, 94-140.

Olshausen, B.A., & Field, D.J. (2004). Sparse coding of sensory inputs. *Curr Opin Neurobiol, 14,* 481-487.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience, 4*, 1244-1252.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2003). Asymmetries in stimulus comparisons by monkey and man. *Current Biology, 13*, 1803-1808.

O'Riordan, M.A., Plaisted, K. (2001). Enhanced Discrimination in Autism. *Quarterly Journal of Experimental Psychology*, *54,* 961-979.

Panis, S., Wagemans, J., & Op de Beeck, H. P. (2010). Dynamic Norm-based Encoding for Unfamiliar Shapes in Human Visual Cortex. *J Cogn Neurosci, 23*, 1829-1843.

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

Pellicano E., Jeffery L., Burr D., & Rhodes G. (2007). Abnormal adaptive face-coding mechanisms in children with autism spectrum disorder. *Curr Biol., 17,* 1508-1512.

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends Cogn Sci., 16,* 504-510.

Plaisted, K. C. (2001). Reduced generalization in autism: An alternative to weak central coherence. In J.A. Burack, T. Charman, N. Yirmiya, & P.R. Zelazo (Eds.), The development of autism: Perspectives from theory and research. (pp. 149-169). New Jersey: Lawrence Erlbaum.

Polk, T. A., Behensky, C., Gonzalez, R., & Smith, E. E. (2002). Rating the similarity of simple perceptual stimuli: Asymmetries induced by manipulating exposure frequency. *Cognition, 82*, B75-88.

Posner, M.I., & Keele, S.W. (1968). On the genesis of abstract ideas. Journal of Experimental Psychology, 77, 353–363.

Renninger, L.W., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research, 44*, 2301-2311.

Rhodes, G., Jeffery, L., Watson, T. L., Clifford, C. W. G., & Nakayama, K. (2003). Fitting the mind to the world: Face adaptation and attractiveness aftereffects. *Psychological Science, 14*, 558-566.

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Res*, 46, 2977–2987.

Rossion, B., Gauthier, I., Tarr, M. J., Despland, P., Bruyer, R., Linotte, S., & Crommelinck, M. (2000). The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: An electrophysiological account of face-specific processes in the human brain. *NeuroReport, 11*, 69-74.
Said, C. P., Dotsch, R., & Todorov, A. (2010a). The amygdala and FFA track both social and non-social face dimensions. *Neuropsychologia, 48,* 3596–3605.

Said, C. P., Moore, C. D., Norman, K. A., Haxby, J. V., & Todorov, A. (2010b). Graded representations of emotional expressions in the left superior temporal sulcus. *Front Syst Neurosci, 4*, 6.

Sams, M., Hietanen, J. K., Hari, R., Ilmoniemi, R. J., & Lounasmaa, O. V. (1997). Face-specific responses from the human inferior occipito-temporal cortex. *Neuroscience, 77*, 49-55.

Schweinberger, S. R., Kloth, N., & Jenkins, R. (2007). Are you looking at me? Neural correlates of gaze adaptation. *NeuroReport, 18*, 693-696.

Schweinberger, S. R., Pickering, E. C., Jentzsch, I., Burton, A. M., & Kaufmann, J. M. (2002). Event-related brain potential evidence for a response of inferior temporal cortex to familiar face repetitions. *Cognitive Brain Research, 14*, 398-409.

Shepard, R. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1*, 54-87.

Sutherland, A, & Crewther, D.P. (2010). Magnocellular visual evoked potential delay with high autism spectrum quotient yields a neural mechanism for altered perception. *Brain, 133,* 2089-2097.

Tanaka, J. W., Curran, T., Porterfield, A. L., & Collins, D. (2006). Activation of preexisting and acquired face representations: The N250 event-related potential as an index of face familiarity. *Journal of Cognitive Neuroscience, 18*, 1488-1497.

Torgerson, W.S. (1965). Multidimensional scaling of similarity. *Psychometrika, 30*, 379-393.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., et al. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Res., 168,* 242–249.

Tsao, D.Y., Freiwald, W.A. (2006). What's so special about the average face? *Trends Cogn Sci, 10,* 391-393.

Tversky., A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q J Exp Psychol A, 43,* 161–204.

van Boxtel, J.J., & Lu, H. (2013). A predictive coding perspective on autism spectrum disorders. *Front Psychol., 4,* 19.

Van de Cruys, S., de-Wit, L., Evers, K., Boets, B., & Wagemans, J. (2013). Weak priors versus overfitting of predictions in autism: Reply to Pellicano and Burr (TICS, 2012). *IPerception, 4,* 95-97.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: predictive coding in autism. *Psychol Rev., 121,* 649-675.

Wark, B., Lundstrom, B.N. & Fairhall, A. (2007). Sensory Adaptation. *Curr Opin Neurobiol* **17**, 423-429.

Wark, B., Fairhall, A. & Rieke, F. (2009). Timescales of inference in visual adaptation. *Neuron 61*, 750-761 (2009).

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature, 428*, 557-561.

Wechsler, D., & Hsiao-pin, C. (2011). WASI-II: Weschler Abbreviated Scale of Intelligence. Pearson.

Weiner, K.S., Sayres, R., Vinberg, J., & Grill-Spector, K. (2010). fMRI-adaptation and category selectivity in human ventral temporal cortex: regional differences across time scales. J *Neurophysiol, 103,* 3349-65.

Worsley, K. J. & Friston, K. J. (1995). Analysis of fMRI Time-Series Revisited - Again. *NeuroImage, 2,* 173-181.