UC Davis

UC Davis Electronic Theses and Dissertations

Title

Understanding Scalp Electrophysiology via Machine Learning: Insights into Working Memory and Credit Assignment

Permalink

https://escholarship.org/uc/item/0x07p1xq

Author

Carrasco, Carlos Daniel

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Understanding Scalp Electrophysiology via Machine Learning: Insights into Working Memory and Credit Assignment

By

CARLOS DANIEL CARRASCO
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Neuroscience

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Steven J. Luck, Co-Chair

Erie D. Boorman, Co-Chair

Andrew Yonelinas

Joy Geng

George Ron Mangun

Committee in Charge

2024

Acknowledgements

Thank you to everyone who took the time to engage in discussion, present work, gave feedback, and was a friend. The last several years have been eye opening not just in the knowledge built throughout my studies, but also in the experiences that have been shared with all of you. Davis has been a safe haven and the Center for Mind and Brain a second home. The neuroscience community as a whole has supported me immensely and the core faculty at the Center for Neuroscience really contributed to the development of my school of thought in early years. I am eternally grateful to Steve and Erie for helping to form my school of thought in the later years of study and onwards. Thank you to Marian Berryhill for giving me the time and space at UNR really affording me the opportunity to attend graduate school. To my friends if you are ever reading this thank you for the good times, may we continue to have them. Thank you to my family who is always in both heart and mind. Wish you all blessings.

General Abstract

Human electrophysiological scalp signals were initially recorded by Hans Berger in 1924. One hundred years later, we have seen an explosion of findings using the electroencephalogram (EEG) to examine neurocognitive processes, and the field continues to push forward with increased use of ever more complex methods. In this dissertation, we first acquaint ourselves with canonical event related potential (ERP) components that have been extensively studied via univariate methods and then discuss how we can leverage machine learning to better detect these components. In the second part of this work, we ask whether a canonical ERP component, the P3b, is related to enhanced working memory representations via behavior, univariate analyses, and decoding of the information content in the neural signals. In the final chapter we venture away from "ERP components" and use a decision-making paradigm and machine learning techniques to ask whether working memory is reactivated to facilitate credit assignment during contingency learning. We applied machine learning to both voltages and alpha band power to determine whether we can decode the sensory features of memories at feedback epochs when items are brought back to mind. This work highlights the effectiveness of machine learning methods in the analysis of noninvasive human electrophysiological data and explores the information content of working memory representations as assessed via neural signals.

Table of Contents

Acknowledgements ii

General Abstract iii

Chapter 1: General Introduction 1

Chapter 2: Using Multivariate Pattern Analysis to Increase Effect Sizes for Event-Related Potential Analyses 20

Abstract 21

1. Introduction 22

2. Methods 29

3. Results 45

4. Discussion 64

Supplementary Materials 76

Chapter 3: Enhanced Working Memory Representations for Rare Events 90

Abstract 91

Significance Statement 92

Introduction 93

Methods 95

Results 103

Discussion 109

Chapter 4: Reinstatement of Working Memory Representations Facilitates Credit Assignment 119

Abstract 120

Significance Statement 121

Introduction 122

Methods 124

Results 139

Discussion 145

Chapter 5: General Discussion 154

Chapter 1: General Introduction

Electrophysiology has enabled researchers to gain a breadth of knowledge in the neurosciences. Initial findings of the electrical nature of nervous system in animals date back to the 18th century (Galvani, 1792). Luigi Galvani, a physician, showed that dissected frog legs could be made to elicit muscular contractions via electrical stimulation of muscle fibers, and he developed a theory of intrinsic animal electricity driving the actions of nerves and muscles. Upon Galvani's death in 1798, there was a divide among the scientific community as to whether the nature of the nervous system was intrinsically electrical or if the finding was an artifact due to dissimilar metals causing electrical conductance through tissue eliciting a response; the latter idea was put forth by prominent physicists (Volta & Banks, 1800). The issue persisted for nearly 50 years, until Emil du Bois-Reymond, a physiologist, cleverly devised both the equipment and the experiments needed to settle the debate. In his seminal work, du Bois-Reymond demonstrated nerve conduction, measured nerve impulses, showed excitation and inhibition of nerve function, and showed differences in electrical properties of nerves and muscles (du Bois-Reymond, 1848). Critical to his work, he developed non-polarized electrodes, the electrometer, and the galvanometer, which allowed him to demonstrate the intrinsic nature of animal electricity (Finkelstein, 2015). More importantly, this work led to the discovery of the action potential, which is the hallmark process by which we understand information transfer among nerve cells. A few decades later, the physiologist Richard Caton would publish his seminal work using electrophysiology to study animal brain electricity, where he witnessed rhythmic electrical activity which he termed brain waves in 1875; this finding went on to influence the creator of the noninvasive brain wave machine in the 20th century (Caton, 1875; Haas, 2003).

While the action potential was discovered in 1848, another debate arose among the physiologists as to the nature of the nervous system. Two predominant theories were *reticular*

theory and what was coined the *neuron doctrine*. The former put forth a connectivist view where the nervous system was composed of a continuous network of neural fibers through which neural impulses propagated information. The latter put forth the notion that nerve cells were discrete units that were supported by astrocytes and other glial cells. While many great scientists contributed to this debate, the two major players were primarily Camillo Golgi, and Santiago Ramon y Cajal. Camillo Golgi developed the black stain or 'Golgi stain' method through which nerve cells could be visualized without causing damage to them (Golgi, 1873) Ramon y Cajal, using this staining method, found evident that neural bodies were not a continuous network, but were instead discrete units with gaps between them (y Cajal, 1888). As a result, Ramon y Cajal's findings were conceptualized as the neuron doctrine by Wilhelm Gottfried von Waldeyer-Hartz, who explicitly named nerve cells 'neurons' and coined the term 'synapse' for the small gap through which transmission of information was made between neurons (Waldeyer, 1891). A little over a decade later, both Golgi and Ramon y Cajal received the Nobel Prize in Physiology or Medicine in 1906 (Fishman, 2007; Grant, 2007).

As physiologists were making great headway in understanding neurophysiology and neuroanatomy, there were two emerging schools of thought coming to fruition within psychology in the late 19th century. Structuralism, whose creation is largely credited to Wilhelm Wundt (Asthana, 2015; Ulrich & Kaup, 2024), focused on understanding conscious experience primarily through introspection and the breakdown of complex mental processes into more basic subcomponents. Wundt went on to establish the first experimental psychology laboratory at the University of Leipzig in 1879 (Bringmann & Ungerer, 1980). There were issues with structuralism, however. Primarily, introspection was too subjective to lead to replicable results. Moreover, structuralism completely focused on conscious processes and failed to consider the

role of unconscious processes on behavior. As a response to structuralism, functionalism emerged and focused on an understanding of the adaptive functions of mental processes and behavior. This view was popularized by psychologist William James (1983). In 1890, James published *The Principles of Psychology*, which treated the study of the mind with rigorous empiricism and pragmatism and focused on studying mental processes to uncover their functions and purpose in everyday life (James, 1890).

A related school of thought in physics and physiology was predominantly founded by Gustav Theodor Fechner in his *Elements of Psychophysics* (Fechner, 1860). Fechner put forth the notion that the relationship between physical stimuli and perceptual experiences could be quantified and their meaning elucidated mathematically. Hermann Von Helmholtz was one of the first individuals to study sensory processes via this discipline, making seminal contributions to the understanding of visual and auditory processes. Helmholtz was meticulous in his study of vision, contributing to theories such as the trichromatic theory, developing tools to study ganglion cells and their relationship to neurons, and examining how unconscious processes influence behavior (Burdon-Sanderson & Hilton, 1903; Patton, 2008; Westheimer, 1983). More importantly, Helmholtz's pragmatism, empiricism, and focus on experimental methods influenced Wundt's approach and the emerging development of the field of experimental psychology (Kantor, 1979; Kim, 2006).

This combination of electrophysiology, a functional understanding of neurons, the development of psychophysics, the creation of falsifiable theories regarding perception, and the development of experimental psychology allowed the scientists of the early 20th Century to make major headway in understanding neuroscience at all levels, ranging from molecules to minds. It was through these means that the field of *psychophysiology*, a term introduced by William James

(1890), would come into full fruition early in the 20th century. This new field was largely sparked by the discovery of electroencephalogram by Hans Berger. Berger was a German physician who was determined to study telepathic communication between humans via long range signals he theorized were created by the brain. At 19 years old during his university studies, he suffered a fall from a horse, where he was severely injured. Coincidentally, his sister hundreds of miles away had a bad feeling about Hans and urged her father to reach out via telegram to determine whether he was alright. This coincidence left a great impression on Hans, and it became his mission to understand this parapsychological phenomenon (Gloor, 1969; Millett, 2001). Having heard of Richard Caton's work, Berger knew that it was possible to record the electrical activity generated by animal brains (Haas, 2003). He theorized that the electrical activity generated by the brain could be picked up noninvasively, and he made some initial attempts to use a galvanometer to pick up this activity. He soon realized that the device was too insensitive to pick up the small signals generated by the brain at the scalp, so he developed his own device which he termed the electroencephalograph. The first EEG recording capturing human brain signals was done on a 17-year-old patient with a postoperative skull defect in 1924 (Gloor, 1969). Berger did not immediately publish his findings, fearing backlash, and sure enough they were met with major criticism when he finally published what he felt like were clearer recordings (Berger, 1929). It took some follow-ups using faraday cages from more established physiologists, some of whom also noted the existence of visually evoked potentials, to validate his work (Adrian & Matthews, 1934a, 1934b; Gibbs et al., 1935; Jasper & Carmichael, 1935). In any case, Berger was vindicated and commemorated for his seminal contribution to the study of neurophysiology. While Berger never reached his goal of demonstrating telepathy, the EEG method he pioneered has been used in both clinical and basic

science work related to brain processes as a method of directly measuring brain activity and investigating its relationship to psychological phenomena.

The 1930s and 1940s were largely marked by the wide dissemination of this new method, although research slowed in the 1940s due to the second world war. The first EEG laboratory was opened at Massachusetts General Hospital in 1936 (Stone & Hughes, 2013). The EEG was used to study topics related to epilepsy, brain trauma, sleep, and mental processes. Of particular interest, early studies of sensory evoked potentials showed that brain waves recorded via the EEG could change when auditory stimuli were presented to participants (Davis, 1939). Somatosensory evoked potentials were discovered in the post-war era, and averaging began to be used to improve the signal-to-noise ratio of the evoked potentials (Dawson, 1954). Several researchers contributed to the methods used to measure and quantify evoked potentials, which ultimately ended up being termed *Event-Related Potentials* (ERPs). The discovery of the *contingent negative variation* (CNV), an ERP component that is characterized by a negative deflection occurring prior to an anticipated stimulus, showed that preparatory neural activity could be detected in the EEG signal. It was this discovery that really sparked the modern era of ERP research, as it was the first study showing that a cognitive process could be directly assessed via non-invasive scalp recordings (Walter et al., 1964).

Emmanuel Donchin, who founded the Cognitive Psychophysiology Laboratory the University of Illinois Urbana Champaign, made several contributions to ERP methods and conceptualization that allowed greater insights into cognitive processes thanwere initially thought possible (Donchin et al., 1978; Empson, 1986; Fabiani et al., 2018). This was primarily incited by the discovery by Sutton et al. (1965) of the P300 component, which is a positive event related potential deflection that changes in size depending on stimulus probability at

approximately 300 ms post stimulus onset. Donchin went on to extensively study the P300 among other components, showing that "Science depends on data that are recorded reliably, analyzed properly, and interpreted creatively," cementing the notion that ERPs could provide meaningful insight into cognitive processes (Fabiani et al., 2018; Picton et al., 2000).

Whereas Donchin's contributions were primarily methodological, Steven Hillyard at the University of California San Diego showed that ERPs could be used to answer questions about neurocognitive function that could not be answered using other methods. In particular, he used ERPs to provide the first concrete evidence that attention influenced sensory processing in audition as early as 100 ms post stimulus onset (Hillyard et al., 1973). Hillyard and colleagues later on extended this work to the visuospatial domain and formulated the sensory gain control theory, which elucidates how selective attention enhances the processing of sensory information to an attended to stimulus providing a basis for distinguishing between separate mechanisms of attention suppression and facilitation (Hillyard et al., 1995, 1998; Luck & Hillyard, 1994; Mangun et al., 1993).

Throughout the remainder of the 20th Century, several ERP components were discovered and used to study various cognitive processes. Of particular importance to the work that will be presented in this thesis are the N170 component, the N400 component, the N2 posterior contralateral (N2pc) component, the lateral readiness potential (LRP), the error related negativity (ERN), the P3b component, and the mismatch negativity (MMN). These voltage deflections are related to specific cognitive processes and have been operationalized as *components* of an ERP waveform. They are further discussed in Chapter 2 which describes methodological work related to the study of these and potentially various other components using multivariate pattern analysis (MPVA), otherwise known as *decoding* techniques.

There are many subfields in the study of neuroscience but given that the work in this thesis is primarily related to human electrical recordings, the experimental work is in the domain of cognitive processes. The research that will be described in Chapters 3 and 4 involve the *working memory* system, a cognitive system that was initially described by Miller et al. (1960) and that was then formalized into a highly influential multicomponent model by Alan Baddeley and Graham Hitch in (1974). The concept of working memory is often distinguished from the older concept short-term memory by emphasizing the processing and manipulation of information. It was later discovered that working memory had neural correlates such as those seen in the seminal work by Patricia Goldman-Rakic and her colleagues, where sustained neural activity in primate prefrontal cortex was observed during delayed match to sample tasks (Funahashi et al., 1990, 1991; Goldman-Rakic, 1991, 1995). Further work with neural recordings and neural imaging have implicated various brain areas engaged in working memory processes such as the parietal cortex, temporal cortex, occipital cortex, thalamus, basal ganglia, and hippocampus (Constantinidis & Klingberg, 2016; D'Esposito & Postle, 2015). There have been several updates since the initial working memory model, and there are various groups with ever-evolving theories as to how the system should be formalized. However, the key insight that persists is that the working memory system facilitates the maintenance and manipulation of information for agents to interact with their environments in goal-driven ways in the moment (Baddeley, 2010; Luck & Vogel, 2013; Ma et al., 2014; E. K. Miller et al., 2018).

A prominent ERP component that has been discussed as being related to working memory is the P3b (Donchin et al., 1975; Polich, 2007, 2012; Squires et al., 1977). In the third chapter of this work, we introduce a new variant of the classic *oddball* paradigm that has been traditionally used to elicit the P3b component. In the new variant of this paradigm, behavioral

probes were used to provide a measure of working memory precision, making it possible to determine whether working memory representations are more precise for P3b-eliciting oddball stimuli. We also performed ERP decoding based on the timepoint-by-timepoint scalp topography, a method which is extensively examined and validated in chapter 2. These decoding approaches are extremely sensitive to real-time cognitive processes, and researchers have used them to decode the information content of ERPs related to different faces, emotions, object orientations, colors, motion and so on (Bae et al., 2023; Bae & Luck, 2019; Bae & Luck, 2018; Daly, 2023; Dowd et al., 2017; Li et al., 2022; Nadra et al., 2023; Noah et al., 2020; Schreiner et al., 2023; Trammel et al., 2023). They in effect allow us to gain insight into the stimulus class encoded by the brain which is used to by an algorithm to decode. The main theoretical question in Chapter 3 is whether working memory representations are improved for rare events. To preview the work, we found both behavioral and ERP decoding evidence of improved working memory representations for rare, P3b-elicitng stimuli, possibly reflecting an increased allocation of attention.

Chapter 4 assesses another possible role of working memory in cognition, namely the process of reactivating quiescent working memory representations to be used online for relevant stimulus/action pairings after the occurrence of a reward, a process that is hypothesized to allow for *credit assignment*, particularly following delayed reinforcement. Credit assignment is the process of determining which actions, decisions or components in a complex system are responsible for specific outcomes. It was originally conceived in the context of machine learning/artificial intelligence, where algorithms were created to give agents the ability to assign credit to useful stimulus/action outcome pairings. This term was made popular after the publication of the seminal textbook, *Reinforcement Learning: An Introduction* by Richard Sutton

and Andrew Barto in 1990 (R. S. Sutton & Barto, 2018). This formalization was predominantly inspired by the contingency learning work conducted by Ivan Pavlov, B.F. Skinner, and Edward Tolman which often is referenced in terms of habit-based or goal directed processes (Dayan & Berridge, 2014; Dayan & Niv, 2008). Pavlov showed that classical conditioning could create behavioral and physiological responses on the basis of stimulus/outcome pairings, in which a bell was "given credit" for the subsequent delivery of food (Pavlov, 1897, 1927). Skinner showed through operant conditioning that reinforcement and reward schedules could elicit a strengthening and weakening of behaviors (Skinner, 1933, 1935, 1965, 2019). Tolman showed that rodents could make inferences about task spaces to gain reward, thus allowing for credit assignment to generalize its use in purposive behavior (Tolman, 1932, 1948, 1949).

Reinforcement learning is often dichotomized as being model-free or model-based (Dayan & Berridge, 2014; Dolan & Dayan, 2013). Briefly, *model-free* reinforcement learning refers to methods that update expected values of actions or states directly and incrementally based on reward prediction errors that are thought to underlie habitual control and simple associative learning processes. *Model-based* methods, on the other hand, use an internal forward model of a task's possible transitions (or task space) to compute expected future values of actions, allowing value functions to be updated immediately following new learning, and is thought to underlie goal-directed control (Doll et al., 2012). Credit assignment can directly reinforce actions taken recently according to eligibility traces which are exponentially decaying representations of a recent activity/state supporting a model-free process (Bogacz et al., 2007). It can also utilize a model of the task structure to assign credit only to specific past choices and to even generalize credit to untaken but associated choices which is supporting model-based processes (Doll et al., 2012; Witkowski et al., 2022). There are various formalizations of what

this ‘task space’ could be or should be called, but the key is that a model of the task structure can facilitate an understanding of which critical components of a task or which past choices to assign credit for outcomes, which is crucial for real-world contingency learning (Jocham et al., 2016; Rushworth et al., 2011; Walton et al., 2010). Credit assignment has been shown to be primarily linked to areas known to be involved in reward, learning, and executive function such as: the orbital frontal cortex, prefrontal cortex, basal ganglia (nucleus accumbens/striatum), amygdala, hippocampus, anterior cingulate cortex, and cerebellum (Lamba et al., 2023; Rolls, 2019; Rosenbloom et al., 2012). The neurotransmitter dopamine appears to be one of the main contributors to the process of credit assignment allowing for the encoding of rewards and facilitation of goal directed behaviors (Hamid et al., 2021; Parker et al., 2022).

The reward system and the working memory system have been implicated in these processes, but how the working memory system is being used to facilitate model-based credit assignment via sustained or reactivated memory representations when outcomes are only delivered after some delay is relatively unexplored. To probe this experimentally, we designed a novel stochastic 2-armed bandit task where both arms had slowly drifting probabilities of reward which participants needed to assign credit to received favorable outcomes and integrate information about counterfactuals (the option that was not chosen). Previewing this work, we found evidence that the representations encoded during visual presentation of stimuli are reinstated at epochs of feedback when the representations are no longer on the screen or decodable in a sustained manner, to facilitate contingency learning.

Neuroscience, an interdisciplinary undertaking, has taken a conglomerate of fields working separately and is continuously refined by researchers unifying them over time towards a common aim to unravel deep insights into brain processes. The discovery of the EEG as was

highlighted in this introduction was merely 100 years ago, and the field of cognitive neuroscience has existed for less than 50 years.; the first issue of the Journal of Cognitive Neuroscience was published in 1968 The discovery of the various ERP components that will be examined in Chapter 2 happened within the last 60 years, and the working memory became a major area of research within the last 50 years. With the continual explosion of computing, we are now able to leverage algorithms in a way that is also causing an exponential growth of knowledge in our field.

References

Adrian, E. D., & Matthews, B. H. C. (1934a). The Berger rhythm: potential changes from the occipital lobes in man. *Brain*, *57*(4), 355–385.

Adrian, E. D., & Matthews, B. H. C. (1934b). The interpretation of potential waves in the cortex. *The Journal of Physiology*, *81*(4), 440.

Asthana, H. S. (2015). Wilhelm wundt. *Psychological Studies*, *60*(2), 244–248. https://doi.org/10.1007/s12646-014-0295-1

Baddeley, A. (2010). Working memory. In *Current Biology* (Vol. 20, Issue 4). https://doi.org/10.1016/j.cub.2009.12.014

Baddeley, A. D., & Hitch, G. J. (1974). Working memory (Vol. 8). *New York: GA Bower (Ed), Recent Advances in Learning and Motivation*.

Bae, G. Y., Leonard, C. J., Hahn, B., Gold, J. M., & Luck, S. J. (2023). Assessing the information content of ERP signals in schizophrenia using multivariate decoding methods. *NeuroImage: Clinical*, *25*. https://doi.org/10.1016/j.nicl.2020.102179

Bae, G. Y., & Luck, S. J. (2019). Decoding motion direction using the topography of sustained ERPs and alpha oscillations. *NeuroImage*, *184*, 242–255. https://doi.org/10.1016/j.neuroimage.2018.09.029

Bae, G.-Y., & Luck, S. J. (2018). Dissociable Decoding of Spatial Attention and Working Memory from EEG Oscillations and Sustained Potentials. *The Journal of Neuroscience*, *38*(2), 2860–17. https://doi.org/10.1523/JNEUROSCI.2860-17.2017

Berger, H. (1929). Über das elektroenkephalogramm des menschen. *Archiv Für Psychiatrie Und Nervenkrankheiten*, *87*(1), 527–570.

Bogacz, R., McClure, S. M., Li, J., Cohen, J. D., & Montague, P. R. (2007). Short-term memory traces for action bias in human reinforcement learning. *Brain Research*, *1153*(1), 111–121. https://doi.org/10.1016/j.brainres.2007.03.057

Bringmann, W. G., & Ungerer, G. A. (1980). The Foundation of the Institute for Experimental Psychology at Leipzig University. In *Psychol. Res* (Vol. 42).

Burdon-Sanderson, J., & Hilton, H. (1903). *Hermann von Helmholtz*. Nature Publishing Group UK London.

Caton, R. (1875). Electrical currents of the brain. *The Journal of Nervous and Mental Disease*, *2*(4), 610.

Constantinidis, C., & Klingberg, T. (2016). The neuroscience of working memory capacity and training. In *Nature Reviews Neuroscience* (Vol. 17, Issue 7, pp. 438–449). Nature Publishing Group. https://doi.org/10.1038/nrn.2016.43

Daly, I. (2023). Neural decoding of music from the EEG. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-022-27361-x

Davis, P. A. (1939). Effects of acoustic stimuli on the waking human brain. *Journal of Neurophysiology*, *2*(6), 494–499.

Dawson, G. D. (1954). A summation technique for the detection of small evoked potentials. *Electroencephalography & Clinical Neurophysiology*.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. In *Cognitive, Affective and Behavioral Neuroscience* (Vol. 14, Issue 2, pp. 473–492). Springer New York LLC. https://doi.org/10.3758/s13415-014-0277-8

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. https://doi.org/10.1016/j.conb.2008.08.003

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, *66*, 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. In *Neuron* (Vol. 80, Issue 2, pp. 312–325). https://doi.org/10.1016/j.neuron.2013.09.007

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. In *Current Opinion in Neurobiology* (Vol. 22, Issue 6, pp. 1075–1081). Elsevier Ltd. https://doi.org/10.1016/j.conb.2012.08.003

Donchin, E., Ritter, W., McCallum, W. C., & others. (1978). Cognitive psychophysiology: The endogenous components of the ERP. *Event-Related Brain Potentials in Man*, *349*, 411.

Donchin, E., Tueting, P., Ritter, W., Kutas, M., & Heffley, E. (1975). ON THE INDEPENDENCE OF THE CNV AND THE P300 COMPONENTS OF THE HUMAN AVERAGED EVOKED POTENTIAL 1. In *Electroencephalography and Clinical Neurophysiology* (Vol. 38).

Dowd, E. W., Pearson, J. M., & Egner, T. (2017). Decoding working memory content from attentional biases. *Psychonomic Bulletin and Review*, *24*(4), 1252–1260. https://doi.org/10.3758/s13423-016-1204-5

du Bois-Reymond, E. (1848). Untersuchungen über thierische Elektricität. *Annalen Der Physik*, *151*(11), 463–464. https://doi.org/https://doi.org/10.1002/andp.18481511120

Empson, J. (1986). Event-related Potentials. In *Human Brainwaves: The Psychological Significance of the Electroencephalogram* (pp. 36–55). Macmillan Education UK. https://doi.org/10.1007/978-1-349-18312-8_3

Fabiani, M., Kutas, M., Miller, G. A., & Coles, M. G. H. (2018). Emanuel Donchin (1935–2018). *Psychophysiology*, *55*(12). https://doi.org/10.1111/psyp.13302

Fechner, G. T. (1860). *Elements of psychophysics*.

Finkelstein, G. (2015). Mechanical neuroscience: Emil du Bois-Reymond's innovations in theory and practice. *Frontiers in Systems Neuroscience*, *9*(September). https://doi.org/10.3389/fnsys.2015.00133

Fishman, R. S. (2007). The nobel prize of 1906. *Archives of Ophthalmology*, *125*(5), 690–694.

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1990). *Visuospatial Coding in Primate Prefrontal Neurons Revealed by Oculomotor Paradigms* (Vol. 63, Issue 4).

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1991). Neuronal Activity Related to Saccadic Eye Movements in the Monkey's Dorsolateral Prefrontal Cortex. In *JOURNALOFNEUROPHYSIOLOGY* (Vol. 65, Issue 6).

Galvani, L. (1792). *De viribus electricitatis in motu musculari commentarius*. Soc. Typogr.

Gibbs, F. A., Davis, H., & Lennox, W. G. (1935). The Electro-Encephalogram in Epilepsy and in Conditions of Impaired Consciousness. *Archives of Neurology & Psychiatry*, *34*, 1133–1148. https://doi.org/10.1001/archneurpsyc.1935.02250240002001

Gloor, P. (1969). Hans Berger on Electroencephalography. *American Journal of EEG Technology*, *9*(1), 1–8. https://doi.org/10.1080/00029238.1969.11080728

Goldman-Rakic, P. S. (1991). Cellular and circuit basis of working memory in prefrontal cortex of nonhuman primates. *Progress in Brain Research*, *85*, 325–336.

Goldman-Rakic, P. S. (1995). Cellular Basis of Working Memory Review. In *Neuron* (Vol. 14).

Golgi, C. (1873). Sulla struttura della sostanza grigia dell cervello', Gazz. *Med Lombarda*, *33*, 224–246.

Grant, G. (2007). How the 1906 Nobel Prize in Physiology or Medicine was shared between Golgi and Cajal. *Brain Research Reviews*, *55*(2), 490–498.

Haas, L. F. (2003). Hans berger (1873–1941), richard caton (1842–1926), and electroencephalography. *Journal of Neurology, Neurosurgery & Psychiatry*, *74*(1), 9.

Hamid, A. A., Frank, M. J., & Moore, C. I. (2021). Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. *Cell*, *184*(10), 2733-2749.e16. https://doi.org/10.1016/j.cell.2021.03.046

Hart, J. (1983). William James and the Harvard Line of Clinical Functionalism. In *Modern Eclectic Therapy: A Functional Orientation to Counseling and Psychotherapy: Including a Twelve-Month Manual for Therapists* (pp. 29–42). Springer US. https://doi.org/10.1007/978-1-4684-1158-4_3

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science*, *182*(4108), 177–180.

Hillyard, S. A., Mangun, G. R., Woldorff, M. G., & Luck, S. J. (1995). *Neural systems mediating selective attention.*

Hillyard, S. A., Vogel, E. K., & Luck, S. J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *353*(1373), 1257–1270.

James, W. (1890). *The Principles Of Psychology Volume I & II*.

Jasper, H. H., & Carmichael, L. (1935). Electrical potentials from the intact human brain. *Science*, *81*(2089), 51–53.

Jocham, G., Brodersen, K. H. H., Constantinescu, A. O. O., Kahn, M. C. C., Ianni, A. M., Walton, M. E. E., Rushworth, M. F. F. S., & Behrens, T. E. E. J. (2016). Reward-Guided Learning with and without Causal Attribution. *Neuron*, *90*(1), 177–190. https://doi.org/10.1016/j.neuron.2016.02.018

Kantor, J. R. (1979). Wundt, experimental psychology and natural science. *Revista Mexicana de Análisis de La Conducta*, *5*(2), 117–129.

Kim, A. (2006). *Wilhelm Maximilian Wundt*.

Lamba, A., Nassar, M. R., & Feldmanhall, O. (2023). Prefrontal cortex state representations shape human credit assignment. *ELife*, *12*. https://doi.org/10.7554/eLife.84888

Li, Y., Zhang, M., Liu, S., & Luo, W. (2022). EEG decoding of multidimensional information from emotional faces. *NeuroImage*, *258*. https://doi.org/10.1016/j.neuroimage.2022.119374

Luck, S. J., & Hillyard, S. A. (1994). Spatial filtering during visual search: evidence from human electrophysiology. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 1000.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. https://doi.org/10.1038/nn.3655

Mangun, G. R., Hillyard, S. A., & Luck, S. J. (1993). *Electrocortical substrates of visual selective attention.*

Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. In *Neuron* (Vol. 100, Issue 2, pp. 463–475). Cell Press. https://doi.org/10.1016/j.neuron.2018.09.023

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior. Henry Holt and Co*.

Millett, D. (2001). Hans Berger: From psychic energy to the EEG. In *Perspectives in Biology and Medicine* (Vol. 44, Issue 4, pp. 522–542). Johns Hopkins University Press. https://doi.org/10.1353/pbm.2001.0070

Nadra, J. G., Bengson, J. J., Morales, A. B., & Mangun, G. R. (2023). Attention without Constraint: Alpha Lateralization in Uncued Willed Attention. *ENeuro*, *10*(6). https://doi.org/10.1523/ENEURO.0258-22.2023

Noah, S., Powell, T., Khodayari, N., Olivan, D., Ding, M., & Mangun, G. R. (2020). Neural Mechanisms of Attentional Control for Objects: Decoding EEG Alpha When Anticipating Faces, Scenes, and Tools. *Journal of Neuroscience*, *40*(25), 4913–4924. https://doi.org/10.1523/JNEUROSCI.2685-19.2020

Parker, N. F., Baidya, A., Cox, J., Haetzel, L. M., Zhukovskaya, A., Murugan, M., Engelhard, B., Goldman, M. S., & Witten, I. B. (2022). Choice-selective sequences dominate in cortical relative to thalamic inputs to NAc to support reinforcement learning. *Cell Reports*, *39*(7). https://doi.org/10.1016/j.celrep.2022.110756

Patton, L. (2008). *Hermann von Helmholtz*.

Pavlov, I. (1897). The Work of the Digestive Glands. *Извѣстія Императорской Академіи Наукъ*, *7*, 411–412.

Pavlov, I. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Miller, G. A., Ritter, W., Ruchkin, D. S., Rugg, M. D., & Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, *37*(2), 127–152. https://doi.org/10.1111/1469-8986.3720127

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Polich, J. (2012). Neuropsychology of P300. *The Oxford Handbook of Event-Related Potential Components*, 159–188.

Rolls, E. T. (2019). The cingulate cortex and limbic systems for emotion, action, and memory. In *Brain Structure and Function* (Vol. 224, Issue 9, pp. 3001–3018). Springer. https://doi.org/10.1007/s00429-019-01945-2

Rosenbloom, M. H., Schmahmann, J. D., & Price, B. H. (2012). The Functional Neuroanatomy of Decision-Making. In *The Journal of Neuropsychiatry and Clinical Neurosciences* (Vol. 24). http://neuro.psychiatryonline.org

Rushworth, M. F. S., Noonan, M. A. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal Cortex and Reward-Guided Learning and Decision-Making. *Neuron*, *70*(6), 1054–1069. https://doi.org/10.1016/j.neuron.2011.05.014

Schreiner, L., Sieghartsleitner, S., Mayr, K., Pretl, H., & Guger, C. (2023). Hand gesture decoding using ultra-high-density EEG. *International IEEE/EMBS Conference on Neural Engineering, NER*, *2023-April*. https://doi.org/10.1109/NER52421.2023.10123901

Skinner, B. F. (1933). The rate of establishment of a discrimination. *Journal of General Psychology*, *9*(2), 302–350. https://doi.org/10.1080/00221309.1933.9920939

Skinner, B. F. (1935). The Generic Nature of the Concepts of Stimulus and Response. *The Journal of General Psychology*, *12*(1), 40–65. https://doi.org/10.1080/00221309.1935.9920087

Skinner, B. F. (1965). *Science and human behavior* (Issue 92904). Simon and Schuster.

Skinner, B. F. (2019). *The behavior of organisms: An experimental analysis*. BF Skinner Foundation.

Squires, K. C., Donchin, E., Herning, R. I., & Mccarthy, G. (1977). ON THE INFLUENCE OF TASK RELEVANCE AND STIMULUS PROBABILITY ON EVENT-RELATED-POTENTIAL COMPONENTS *. In *Electroencephalography and Clinical Neurophysiology* (Vol. 42).

Stone, J. L., & Hughes, J. R. (2013). Early history of electroencephalography and establishment of the American Clinical Neurophysiology Society. *Journal of Clinical Neurophysiology*, *30*(1), 28–44.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, *150*(3700), 1187–1188.

Tolman, E. C. (1932). *Purposive behavior in animals and men*. Univ of California Press.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189.

Tolman, E. C. (1949). There is more than one kind of learning. *Psychological Review*, *56*(3), 144.

Trammel, T., Khodayari, N., Luck, S. J., Traxler, M. J., & Swaab, T. Y. (2023). Decoding semantic relatedness and prediction from EEG: A classification method comparison. *NeuroImage*, *277*, 120268. https://doi.org/10.1016/j.neuroimage.2023.120268

Ulrich, R., & Kaup, B. (2024). Wilhelm Wundt: His bumpy start in science at the University of Tübingen. *History of Psychology*, *27*(1), 90–96. https://doi.org/10.1037/hop0000251

Volta, A., & Banks, J. (1800). I. On the electricity excited by the mere contact of conducting substances of different kinds. *The Philosophical Magazine*, *7*(28), 289–311. https://doi.org/10.1080/14786440008562590

Waldeyer, W. von. (1891). Ueber einige neuere Forschungen im Gebiete der Anatomie des Centralnervensystems1. *DMW-Deutsche Medizinische Wochenschrift*, *17*(44), 1213–1218.

Walter, W. G., Cooper, R., Aldridge, V. J., McCallum, W. C., & Winter, A. L. (1964). Contingent negative variation: an electric sign of sensori-motor association and expectancy in the human brain. *Nature*, *203*(4943), 380–384.

Walton, M. E., Behrens, T. E. J., Buckley, M. J., Rudebeck, P. H., & Rushworth, M. F. S. (2010). Separable Learning Systems in the Macaque Brain and the Role of Orbitofrontal Cortex in Contingent Learning. *Neuron*, *65*(6), 927–939. https://doi.org/10.1016/j.neuron.2010.02.027

Westheimer, G. (1983). Hermann Helmholtz and origins of sensory physiology. *Trends in Neurosciences*, *6*, 5–9.

Witkowski, P. P., Park, S. A., & Boorman, E. D. (2022). Neural mechanisms of credit assignment for inferred relationships in a structured world. *Neuron*, *110*(16), 2680-2690.e9. https://doi.org/10.1016/j.neuron.2022.05.021

y Cajal, S. R. (1888). *Estructura de los centros nerviosos de las aves*.

Chapter 2: Using Multivariate Pattern Analysis to Increase Effect Sizes for Event-Related Potential Analyses

Carlos Daniel Carrasco, Brett Bahle, Aaron Matthew Simmons, Steven J. Luck

Center for Mind & Brain, University of California, Davis

*This work has now been published in the journal Psychophysiology:
https://doi.org/10.1111/psyp.14570

CC-By Attribution 4.0 International, https://creativecommons.org/licenses/by/4.0/.

Abstract

Multivariate pattern analysis approaches can be applied to the topographic distribution of event-related potential (ERP) signals to 'decode' subtly different stimulus classes, such as different faces or different orientations. These approaches are extremely sensitive, and it seems possible that they could also be used to increase effect sizes and statistical power in traditional paradigms that ask whether an ERP component differs in amplitude across conditions. To assess this possibility, we leveraged the open-source ERP CORE dataset and compared the effect sizes resulting from conventional univariate analyses of mean amplitude with two multivariate pattern analysis approaches (support vector machine decoding and the cross-validated Mahalanobis distance, both of which are easy to compute using open-source software). We assessed these approaches across seven widely studied ERP components (N170, N400, N2pc, P3b, lateral readiness potential, error related negativity, and mismatch negativity). Across all components, we found that multivariate approaches yielded effect sizes that were as large or larger than the effect sizes produced by univariate approaches. These results indicate that researchers could obtain larger effect sizes, and therefore greater statistical power, by using multivariate analysis of topographic voltage patterns instead of traditional univariate analyses in many ERP studies.

Keywords: ERPs, EEG, Decoding, Classification, Support Vector Machine, Cross-Validated Mahalanobis Distance

1. Introduction

For decades, ERP researchers have assessed differences in amplitude between experimental conditions with univariate analyses. In these analyses, the mean or peak voltage within a time period is measured for each condition at one or more electrode sites, and univariate statistical analyses (e.g., t tests, analyses of variance) are used to determine whether the differences in amplitude across conditions or groups are statistically significant. Researchers often record from >20 electrodes, but the analyses typically focus on either one electrode or a cluster of electrodes where the component of interest is largest(Zhang & Kappenman, 2023). Researchers sometimes put separate measures from each electrode into a univariate statistical analysis and look for condition x electrode site interactions, but these types of analyses typically have low statistical power (Luck, 2014). Mass univariate analyses can also be used to identify electrode sites showing significant differences between conditions (Maris & Oostenveld, 2007), but they may also suffer from reduced power (Groppe et al., 2011). A major shortcoming of these univariate approaches in general is that they assume that all participants have approximately the same scalp distribution for a given component, which is typically not a good assumption given the large individual differences in cortical folding patterns (Borrell, 2018; Fernández et al., 2016; Kroenke & Bayly, 2018), which can produce substantial differences in scalp topography (Hajizadeh et al., 2021).

To address these shortcomings, researchers have begun to apply multivariate pattern analysis (MVPA) techniques to ERP data to gain insight into brain activity that may not be related to traditional ERPs (Ashton et al., 2022; Chan et al., 2011; Daly, 2023; de Vries et al., 2019; Grootswagers et al., 2017; Gurariy et al., 2022; Hubbard et al., 2019; D. Li et al., 2023; Murphy et al., 2011; Wang et al., 2012; Yang et al., 2021). In particular, machine learning

algorithms can be used to distinguish slight differences between different stimulus classes in the pattern of voltage across electrode sites[1]. This makes it possible for a given algorithm to correctly guess the stimulus class that produced an ERP waveform from new data that were not used for training. In other words, the stimulus class is encoded by the brain and can be decoded by the algorithm. For example, MVPA approaches have been used to decode the identity and emotional expression of a face being perceived (Bae, 2021; Dobs et al., 2019; Nemrodov et al., 2018; Smith & Smith, 2019). Decoding can also be used to gain insight into a range of cognitive domains including motion perception, attention, and working memory (Bae & Luck, 2019; Bae, 2021; Bae & Luck, 2018; Ester & Weese, 2023; Y. Li et al., 2022; Nadra et al., 2023; Noah et al., 2020).

These approaches are extremely sensitive, capable of reliably distinguishing ERP patterns for different faces, different directions of motion, different orientations, and so on. It is therefore plausible that they could also be used to increase power for conventional ERP analyses.

The most common ERP decoding approach uses support vector machines (SVMs). As an example, we will briefly explain how this approach works for decoding whether a participant was shown a face stimulus or a car stimulus, and we will then describe how it can be used to answer traditional ERP questions (e.g., whether the N170 component is larger for faces than for cars). Figure 1a shows a simplified example of SVM decoding in which simulated single-trial data from only two electrode sites (e.g., PO8 and Cz) are used to decode whether the subject had been shown a picture of a face or a picture of a car. Each dot in the scatterplot shows the voltage

[1] Here, we focus on different classes of stimuli, but these methods can also be applied to classes based on other factors, such as responses (e.g., left-hand vs. right-hand responses). The only restriction is that the methods discussed here require discrete classes. Regression-based methods can be used to decode continuous variables (e.g., Collins & Frank, 2018).

from this pair of electrode sites for a given trial. Decoding is usually performed separately at each latency to take advantage of the ERP technique's temporal resolution, so the data in Figure 1a would be from a specific latency (e.g., 170 ms after stimulus onset). Although faces and cars produce similar distributions of voltage at each electrode when considered individually, the SVM is able to find a decision line that can separate the face and car data. Note that this example greatly exaggerates the difference between conditions for the sake of clarity.

To avoid capitalizing on noise (i.e., overfitting), the SVM finds the decision line using a set of training cases and then is tested on cases that were not used for training (indicated by the squares in Figure 1a). A given test case is classified according to which side of the decision line it falls on, and this classification is then compared with the true class to determine whether the classification was correct. The proportion correct for the test cases is then computed to quantify decoding accuracy.

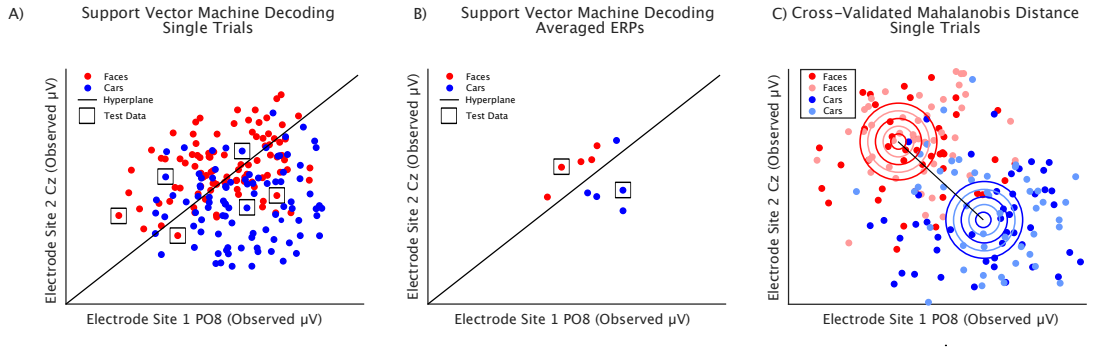


Figure 1. Simplified example of support vector machine (SVM) decoding and the cross-validated Mahalanobis distance (crossnobis distance). This example uses simulated data from only two electrode sites, which can be represented as a 2-dimensional space. The data are shown from a single latency. A) Single-trial decoding. Each dot represents the voltage at these two electrodes on a single trial. The SVM finds a decision line (hyperplane) that best divides the trials into the two classes. Once the SVM is trained, it is tested with data from trials that were not used to train the SVM (shown as dots surrounded by squares). The SVM classifies these test cases according to which side of the decision line they fall on. The proportion of these classifications that were correct (i.e., matched the true classes for these trials) is then used to quantify the accuracy of the decoder. B) Decoding from averaged ERPs. The 80 single trials for each stimulus class in the original dataset are divided into five sets of 16 trials and then averaged to create five averaged ERPs for each class. Four of the averaged ERPs from each class are used for training and

the remaining averaged ERP from each class is used for testing. C) Crossnobis algorithm. This approach operates on single-trial data. It finds the centroid and multidimensional standard deviation for each class and computes the distance between the two means scaled by the standard deviations. The trials from each class are randomly divided into two subsets to allow for a cross-validation procedure. Here we are showing simulated single-trial data with centroids that are much more distant than would typically be observed in real data.

Of course, we typically have far more than two electrode sites in a typical ERP study. All N sites are used simultaneously for decoding, giving us an N-dimensional space. If we have 30 electrode sites, for example, we will have a 30-dimensional space, and the SVM will determine the optimal hyperplane in this 30-dimensional space for distinguishing between the stimulus classes[2]. A separate SVM is trained and tested at each time point in the ERP waveform, yielding a waveform showing decoding accuracy at each time point (which is roughly analogous to an ERP difference wave because it shows the ability of the decoder to differentiate between the conditions).

Although the training and testing cases shown in Figure 1a are individual trials, single-trial EEG epochs are typically too noisy to allow good decoding of subtle stimulus classes. Instead, the trials for each class are typically divided into multiple sets of 10-20 trials, and an averaged ERP waveform is created from each of these sets. For example, if there were 80 face trials and 80 car trials, we might create five face averages and five car averages, each with 16 trials. This is illustrated in Figure 1b. We would train the SVM on the data from four face averages and four car averages and then test the SVM on the data from the one face average and the one car average that were not used for training (separately at each latency). This yields a smaller number of training and testing cases. However, SVMs are particularly well suited for learning to classify

[2] It is possible to apply a dimensionality reduction method such as principal component analysis to the data prior to decoding, but we have not found much advantage to this. It is also possible to use a nonlinear hyperplane to distinguish between the classes. However, when an SVM is applied to ERP data in which each dimension represents the same type of variable (voltage) and a large number of dimensions (electrodes) is used, we find that a linear hyperplane typically works best.

data on the basis of a small number of training cases, and the increase in signal-to-noise ratio produced by averaging outweighs the reduction in the number of training cases(Gholami & Fakhari, 2017; Grootswagers et al., 2017).

Many conventional ERP experiments also involve comparing two classes of trials[3]. For example, mismatch negativity (MMN) and P3b experiments typically involve comparing standard stimuli with oddball stimuli; N400 experiments typically involve comparing words that vary in terms of a factor such as semantic relatedness or concreteness; and error-related negativity experiments typically involve comparing trials with correct versus incorrect responses. In these experiments, an SVM can also be used to decode which of the two trial types occurred, and decoding accuracy is roughly analogous to the difference in voltage between the two trial types at a given latency.

Just as the difference in voltage is commonly averaged across a measurement window in conventional ERP analyses, decoding accuracy can be averaged across a measurement window. Decoding accuracy can therefore be used as the dependent variable in statistical analyses in exactly the same way as differences in voltage across experimental conditions (although the interpretation is subtly different, as will be described in the Discussion). However, because SVMs use the entire pattern of voltage over the scalp in an optimal manner for each individual participant, it is plausible that they might provide greater sensitivity, larger effect sizes, and improved statistical power relative to simple voltage differences.

[3] In this paper, we will focus on binary classifications in which there are two classes being distinguished (e.g., faces versus cars). However, it is straightforward to extend the approaches discussed here to multiclass problems in which three or more classes are being distinguished (e.g., which of 10 faces is being perceived). This can be achieved by decoding each possible pair and averaging the accuracy across all pairs. However, we have found better results using an approach called *error-correcting output codes* (Dietterich & Bakiri, 1995), in which a set of decoders is combined to achieve a single decision for each test case (Bae & Luck, 2018; Bae, 2019; Bae & Luck, 2021).

Decoding accuracy is computed by taking a categorical variable (correct or incorrect) and aggregating across test cases to create a continuous variable (proportion correct). This is much like calculating proportion correct in a perception experiment. Unfortunately, such measures can have poor psychometric properties, such as an insensitivity to differences near the floor or ceiling. An alternative approach in perception experiments is to apply signal detection theory and compute the d' metric of sensitivity, which represents the difference between the average percept produced by the two classes scaled by the within-class variability of the percept. In decoding, the analogous metric is the Mahalanobis distance, which is illustrated in an idealized manner in Figure 1c. This metric represents the distance between the centroids of the two single-trial distributions, scaled by the pooled multidimensional standard deviation and has previously been used in EEG research to decode differences in visual working memory states (Wolff et al., 2015). The Mahalanobis distance is also analogous to the Cohen's d metric of effect size but extended to a multidimensional space.

The Mahalanobis distance has no maximum value, so ceiling effects are not an issue. However, whereas the chance value for decoding accuracy is usually very clear (e.g., 0.5 for a binary classification), chance is not well defined for the Mahalanobis distance (because distances can never be negative, and noise alone will cause the Mahalanobis distance to be greater than zero). This shortcoming can be solved with a cross-validation approach in which the data are split into random halves (see Method for more details). The result is the cross-validated Mahalanobis distance (or crossnobis distance; Walther et al., 2016).

The goal of the present study was to assess whether SVM-based decoding accuracy and the crossnobis distance might offer advantages for ERP studies in which the goal is to compare the amplitude of an ERP component across two conditions. Given that decoding can reliably

determine which of two faces is being perceived or which of two orientations is being maintained in working memory, perhaps it could provide a more powerful way of asking whether, for example, the N170 component is larger for faces than for cars. To draw a conclusion that is relatively general, we analyzed data for seven different ERP components (N170, N2pc, P3b, N400, mismatch negativity, lateralized readiness potential, and error-related negativity) taken from the ERP CORE dataset (Kappenman et al., 2021). Each component in this dataset is isolated by examining the difference between two conditions (e.g., faces versus cars for the N170 component, error trials versus correct trials for the error-related negativity). In the standard univariate approach, the difference in amplitude between the two conditions is quantified at single electrode site or cluster of electrode sites. We compared this approach with SVM decoding accuracy and the crossnobis distance to see whether the multivariate approaches would be more sensitive to the difference between conditions. Other decoding approaches are also available, but a recent study found that SVM outperformed other methods in the context of the N400 component (Trammel et al., 2023).

The three different analysis methods we examined are on different scales (number of microvolts for the univariate analysis, proportion correct for SVM decoding, and multivariate distance in microvolts for crossnobis distance). To put them on a common scale so that we could make comparisons among them, we converted the set of values for the entire group of 40 participants to the Cohen's d_z metric of effect size (Cohen, 1988). This metric represents the size of an effect as the mean across participants minus chance scaled by the standard deviation across participants. Cohen's d_z is directly related to the statistical power for detecting a difference between experimental conditions in a within-subjects analysis, so a larger d_z is preferable (all else being equal). Converting the results into d_z values made it possible to compare the different

analytic approaches on the same scale and allows for planning of future studies by other researchers (Lakens, 2013). Our primary question, therefore, was whether one or both multivariate techniques yields a larger d_z than the traditional univariate approach.

All the decoding analyses reported here were carried out using tools that are built into versions 10 and higher of ERPLAB Toolbox (Lopez-Calderon & Luck, 2014). These tools allow the user to perform decoding using a graphical user interface without any programming (although we used scripts to automate the process and provide a record of exactly how the data were analyzed). Thus, it is now easy and inexpensive for anyone to apply decoding methods to their own ERP data.

2. Methods

The ERP CORE (Kappenman et al., 2021) consists of stimulus presentation scripts, data, and data analysis scripts for six different experiments that can be used to isolate seven different ERP components. A complete description of the participants, stimuli, tasks, and preprocessing for the ERP CORE experiments can be found in the original paper. Here we provide a brief summary, along with a detailed description of the analyses that were performed for the present study.

2.2 Participants

Forty neurotypical young adults who provided informed consent performed each of the six ERP CORE experiments in a single session. For any given component, participants with EEG artifacts on a large number of trials (> 25% of epochs) were excluded. In addition, one participant had only two error trials for the ERN component and was excluded. These procedures led to the exclusion of 1 – 6 participants per component. Table 1 lists the final sample size for each component. The study was approved by the UC Davis Institutional Review Board.

2.3 Overview of Stimuli and Tasks

Figure 2 illustrates the six experimental paradigms. In the N170 task (Rossion & Jacques, 2011), participants were shown faces, scrambled faces, cars, and scrambled cars ($p = .25$ for each category); they were instructed to indicate whether a given stimulus was intact or scrambled. In the N400 task, (Kutas & Federmeier, 2011), each trial consisted of a prime word followed by a semantically related or unrelated target word; participants were instructed to indicate whether each target word was related or unrelated to the preceding prime word. In the N2pc task, (Kiss et al., 2008), each stimulus array included a blue square, a pink square, and several black squares; participants were instructed to attend to the blue squares in some trial blocks and the pink squares in other trial blocks and indicate whether the gap on the attended-color square in each array was on the top or bottom of that square. In the flankers task used to elicit the lateralized readiness potential (LRP) and error-related negativity (ERN) (Eimer & Coles, 2003; Gehring et al., 2012) each display contained a left- or right-pointing central arrow flanked by distractor arrows that were pointing in the same direction or the opposite direction (with equal probability); participants were instructed to indicate the direction of the central arrow. To achieve an appropriate number of error trials for examining the ERN, feedback was used to encourage fast response times that produced incorrect responses on 10–20% of trials. In the P3b task, (Polich, 2012), the letters A, B, C, D, and E were presented in random order ($p = .2$ each); one letter was defined as the target at the beginning of each block, and participants were instructed to indicate whether each stimulus was the target letter ($p = .2$) or any of the nontarget letters ($p = .8$ for the nontarget category). In the MMN task (Näätänen et al., 2007), participants heard a randomized sequence of standard tones (1000 Hz, 80 dB, $p = .8$) and deviant tones (1000 Hz, 70 dB, $p = .2$); participants were instructed to ignore these tones and watch a silent video.

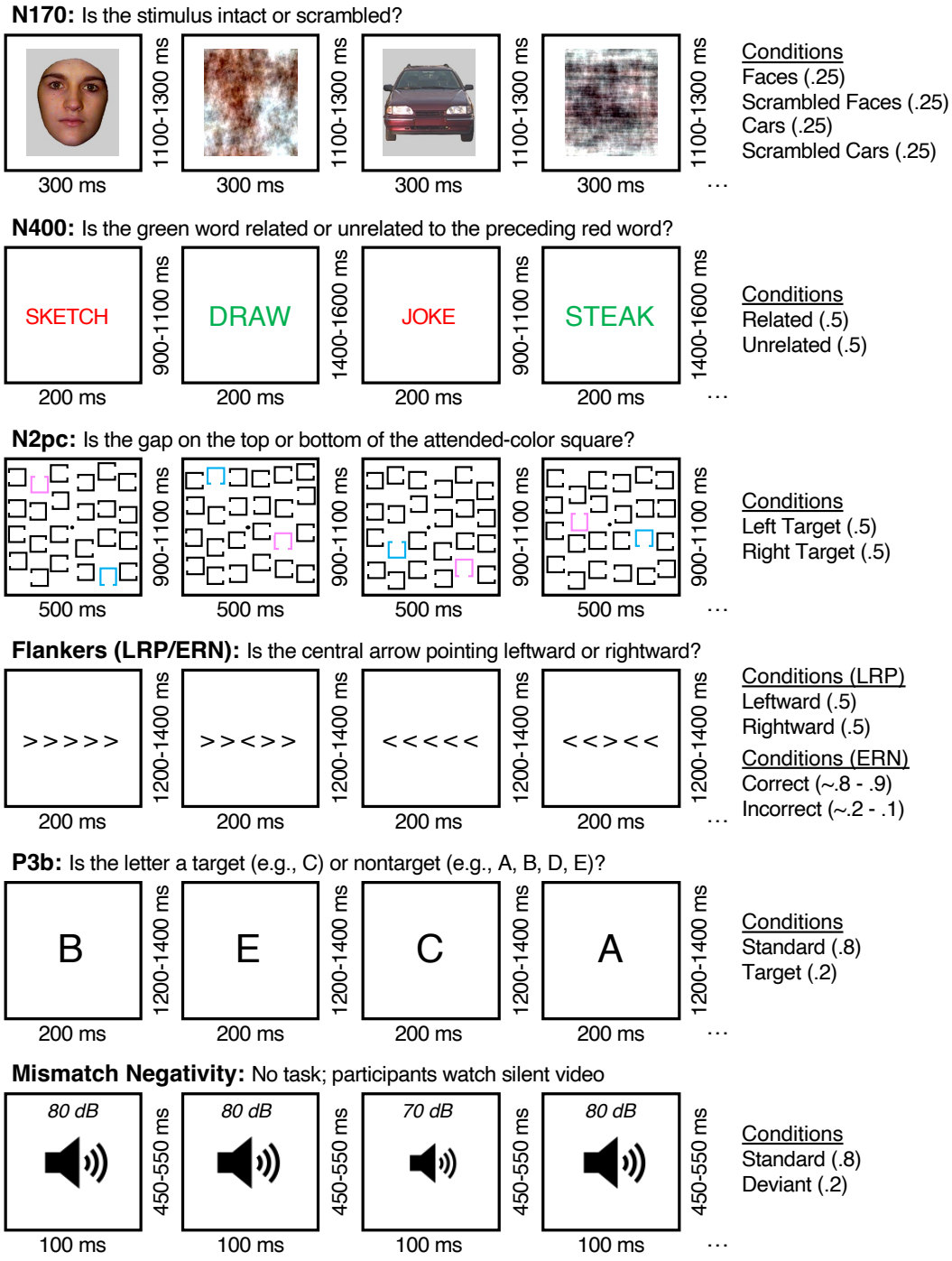


Figure 2. Overview of the ERP CORE tasks. In the N170 task, participants were presented with faces, scrambled faces, cars, and scrambled cars (p = .25 each) and were required to indicate whether each stimulus was intact or scrambled. In the N400 task, participants judged the semantic relatedness of each green target word to the preceding red prime word. The N2pc task involved attending to blue or pink squares in an array of black squares and indicating the location of the gap on the attended-color square. The flankers task was used to isolate both the lateralized readiness potential (LRP) and the error-related

negativity (ERN). Participants were required to indicate the direction of a central arrow flanked by distractors. In the P3b task, participants were presented with a sequence of randomly ordered letters from the set {A, B, C, D, E} and were required to indicate whether each letter was a predefined target letter or one of the four nontarget letters. In the mismatch negativity (MMN) task, participants watched a silent video and were instructed to ignore a sequence of tones that included rare deviant tones (70 dB) and frequent standard tones (80 dB).

2.4 EEG Recording and Processing Procedures

All analyses described here began with the preprocessed, epoched data provided in the ERP CORE resource (Kappenman et al., 2021), which are freely available at https://doi.org/10.18115/D5JW4R. Continuous EEG recordings were obtained using a Biosemi ActiveTwo recording system with active electrodes placed at 30 scalp sites: FP1, F3, F7, FC3, C3, C5, P3, P7, P9, PO7, PO3, O1, Oz, Pz, CPz, FP2, Fz, F4, F8, FC4, FCz, Cz, C4, C6, P4, P8, P10, PO8, PO4, O2. Additional electrodes were placed lateral to the eyes and under the left eye to record the horizontal and vertical electrooculogram (HEOG and VEOG).

Ocular artifacts were corrected offline using independent component analysis, and trials were marked for exclusion if they contained behavioral errors, large voltage deflections in any channel, or ocular artifacts at the time of the stimulus. The data were high-pass filtered (non-causal Butterworth filter, half-amplitude cutoff at 0.1 Hz, 12 dB/octave), downsampled to 256 Hz, and segmented into 1000 ms epochs with 200 ms baselines. The data were referenced offline to the average of the P9 and P10 electrodes (near the left and right mastoids), except that the N170 data were referenced to the average of all scalp sites as is common in N170 studies[4]. See Kappenman et al. (2021) for extended details on the recording methods and preprocessing pipeline.

[4] The location of the reference site(s) has little or no impact on the decoding methods used here.

For the present analyses, we also applied a low-pass filter (non-causal Butterworth filter, half-amplitude cutoff at 20 Hz, 12 dB/octave) to the epoched EEG data prior to all further analyses which is commonly applied prior to measuring components. All preprocessing was conducted in MATLAB using the EEGLAB and ERPLAB toolboxes (Delorme & Makeig, 2004; Lopez-Calderon & Luck, 2014; MathWorks, 2023).

2.5 Traditional Univariate Analyses

The traditional univariate analysis began by computing a difference wave between the key conditions for a given ERP component for each participant (e.g., oddballs versus standards for the P3b component). We then averaged the difference waves across an optimal cluster of electrode sites for that component, as determined by Zhang & Kappenman (2023). An alternative approach would have been to use the maximal electrode site where a component is known to be represented (e.g., site Pz for the P3b) but when we did these univariate analyses, we found that using a cluster of electrodes led to very similar results (see supplementary Figures S2-S4). We therefore chose to use clusters of electrode sites as Zhang & Kappenman (2023) showed cluster-based measurements are as good and often better than measuring from single sites. We then measured the mean voltage across an a priori measurement window and used a one-sample t test to determine whether this difference was significantly different from zero (which is equivalent to asking whether the two conditions were significantly different from each other). Given that the direction of the effects in these paradigms is already known from hundreds of previous studies for a given component, these tests were one-tailed. The alpha for these and all other statistical tests was 0.05. The measurement windows and electrode sites for each component are shown in Table 1. Supplementary Table S1 shows the analogous results when the data were obtained from the single electrode where the component is largest.

The effect size for the mean amplitude across the measurement window or for an individual time point was quantified with Cohen's d_z. This was computed from the voltage in the difference waves (e.g., faces minus cars for N170, contralateral minus ipsilateral for N2pc and LRP), which is identical to computing the effect size comparing the parent waves. For difference scores such as these, the Cohen's d_z equation is:

$$d_z = \left(\overline{X_1} - chance\right) \div \sigma(X_1)$$

where X_1 is equal to the mean difference wave voltage across participants, chance is set to 0 meaning no difference between conditions, and $\sigma(X_1)$ is the standard deviation of the difference wave voltages across participants (see Lakens, 2013).

We also conducted a mass univariate analysis (Groppe et al., 2011; Maris & Oostenveld, 2007) in which we performed a one-sample t test comparing the difference score to zero at each time point between 0 ms and the end of the epoch, using the false discovery rate (FDR) correction for multiple comparisons (Benjamini & Yekutieli, 2001). Given that the direction of the effects in these paradigms is already known from hundreds of previous studies for a given component, one-tailed tests were performed (as shown with gray shading in the waveform figures). However, the direction of these one-tailed tests was based on a specific component that occurs during a particular time window (e.g., the N170 time window in our comparison of faces versus cars). Effects of the opposite polarity were sometimes obtained outside this time window (e.g., a positivity that followed the N170), which were not significant by definition with the one-tailed test. We therefore also conducted two-tailed statistical tests (show with horizontal black lines in the waveform figures).

Because Cohen's d_z is negative when the voltage is more negative in one condition than in another, we took the absolute value of the effect size. The standard error and 95% confidence interval of each d_z value was estimated using bootstrapping, as described in Section 2.8.

Table 1. Univariate analysis parameters for each component, along with the corresponding t test.

Univariate Cluster Parameters

ERP Component	Measurement Window	Electrode Cluster	N	df	Mean Amp t	Mean Amp P	Time Locking Point
N170: Faces – Cars	110 to 150	P7, P8, PO7, PO8, O1, O2	37	36	9.95	p < 0.0001	stimulus onset
N170: Faces - Scrambled Faces	110 to 150	P7, P8, PO7, PO8, O1, O2	37	36	8.62	p < 0.0001	stimulus onset
N170: Cars - Scrambled Cars	110 to 150	P7, P8, PO7, PO8, O1, O2	37	36	2.58	p = 0.014	stimulus onset
N170: Scrambled Faces – Scrambled Cars	110 to 150	P7, P8, PO7, PO8, O1, O2	37	36	3.14	p = 0.003	stimulus onset
N400	300 to 500	Cz, Pz, C3, C4, P3, P4	37	36	13.53	p < 0.0001	stimulus onset
N2pc	200 to 275	P7/8, PO7/8, PO3/4, O1/2, P3/4	35	34	6.90	p < 0.0001	stimulus onset
LRP	-100 to 0	F3/4, FC3/4, C3/4, C5/6, P3/4	37	36	19.90	p < 0.0001	response onset
ERN	0 to 100	Fz, FCz, Cz, FC3, FC4	36	35	9.30	p < 0.0001	response onset
P3	300 to 600	Cz, CPz, Pz, P3, P4	34	33	11.09	p < 0.0001	stimulus onset
MMN	125 to 225	Fz, FCz, Cz, FC3, FC4	39	38	10.3	p < 0.0001	stimulus onset

2.6 Support Vector Machine (SVM) Decoding Analyses

We performed SVM decoding using the decoding tools built into ERPLAB Toolbox, which rely on Matlab's fitcsvm() function. For each component, the two conditions for that component

served as the classes being decoded (e.g., faces vs. cars for the prototypical N170 comparison). Decoding was conducted separately at each time point for each participant. Thus, we obtained a waveform for each participant that reflected the decoder's ability to distinguish between the two conditions at each time point, which is the MVPA analog of the difference in voltage between the two conditions at each time point (i.e., the difference wave).

The SVM analyses used an N-fold cross-validation procedure. Specifically, we randomly divided the data for a given condition into N sets of trials and created an average ERP from each of those N sets of trials. For a given component, we selected the largest N that would yield 10–20 trials per average. For example, there were 80 face trials and 80 car trials for the N170, and we used five sets of averages (five crossfolds) for each condition, with 16 trials per average prior to the exclusion of trials with artifacts or errors. Table 2 shows the number of crossfolds and the minimum number of trials remaining in each condition after exclusion for each component.

Because spurious above-chance decoding accuracy is likely if the number of trials differs between the conditions being decoded for a given participant, we equated (floored) the number of trials used for decoding in each participant. That is, we determined which condition had the smallest number of trials for a given participant after exclusion of trials with errors and artifacts, and we randomly selected an equivalent number of trials for the other condition for that participant. This was done separately for each component. For components with very different numbers of trials per condition (P3b, MMN, and ERN), this meant that a relatively small subset of the trials was selected from the condition with the larger number of trials. For example, there were 160 frequent trials and 40 rare trials for the P3b component, and we used only 40 of the frequent trials (or even fewer given the exclusion of trials with errors and artifacts)[5].

[5] This subsampling procedure is also widely used in conventional univariate analyses comparing peak amplitudes, because peak values are biased to be larger when the waveform is noisier. However, subsampling is not necessary

We performed decoding on the N averaged ERP waveforms for each condition, separately at each time point. We included only the 28 scalp EEG channels, excluding artifact channels and reference channels. The decoding was performed once for each of the N crossfolds, using N-1 of the averaged ERPs from each condition for training and the remaining 1 averaged ERP for each condition for testing. For each of these crossfolds, the SVM was trained using the pattern of voltages across the 28 electrode sites at the time point of interest from each of the N-1 averages for each of the two conditions (e.g., from 4 averages of face trials and 4 averages of car trials). From these training examples, the SVM determined an optimal classification hyperplane for that time point. The SVM was then tested by being given the pattern of voltage across electrodes at that time point for each of the two averages that were not used for training (e.g., 1 average of face trials and 1 average of car trials) and asked to guess the condition for each of these patterns (e.g., face or car). Thus, for a given crossfold, the SVM made two binary guesses, one for each condition. Because there were 2 test cases and therefore 2 guesses for each crossfold, there were 2N test cases and 2N guesses across the N crossfolds (e.g., 10 total guesses across 5 crossfolds for the N170 faces/cars component). This process was then repeated for each time point.

To increase the resolution of the decoding accuracy, we repeated this process 100 times for each participant (100 iterations). For each iteration, we re-randomized the assignment of trials to averages and trained new SVMs. This gave us 200N guesses for a given component at each time point for each fold (e.g., 1000 total guesses for the 5-fold N170 analysis). Decoding accuracy for a given time point was computed as the proportion of guesses that were correct

for mean amplitudes, which are not consistently larger for noisier data (Luck, 2014). Our goal in the present study was to compare the standard univariate approach (which does not require subsampling for mean amplitude) with the standard SVM approach (which does require subsampling).

across these guesses. Because there were two conditions for a given component, and the number of training cases was always the same for the two conditions, chance decoding accuracy was 0.5 at each time point.

For a given component, this process ultimately yielded a decoding accuracy value at each time point for each participant. To determine whether decoding accuracy at each time point was significantly above chance for the group of participants, we conducted a mass univariate analysis using a separate one-sample t test against chance (0.5) at each time point, with the FDR correction for multiple comparisons (Benjamini & Yekutieli, 2001). Although we computed decoding accuracy for each point in the entire epoch, we excluded the prestimulus period from the FDR correction. One-tailed tests were used because below-chance accuracy is not meaningful with this procedure[6]. We also computed the Cohen's d_z metric of effect size:

$$d_z = \left(\overline{X_1} - chance\right) \div \sigma(X_1)$$

where X_1 is equal to the mean of the decoding accuracy values across participants, chance is 0.5 or 50%, and $\sigma(X_1)$ is the standard deviation of the decoding accuracy values across participants. This is exactly the same d_z formula used for the univariate analyses. As detailed in Section 2.8, bootstrapping was used to compute standard errors and 95% confidence intervals for d_z.

In many ERP studies, the amplitude of a component is quantified as the mean voltage within a specific time window. To conduct an analogous analysis with the decoding data, we simply averaged the decoding accuracy values across the time points within the measurement window listed in Table 2 for a given component (e.g., 110-150 ms for Face vs Cars). A different

[6] In the univariate analysis, the voltage at a given point in the difference wave might be significantly greater than zero or significantly less than zero, so a two-tailed test was necessary. For decoding, either a positive or a negative difference would yield an above-chance decoding accuracy value, so a one-tailed test was appropriate. We are not comparing significance values between the univariate and decoding methods, so this does not give decoding an unfair advantage.

approach would be to average the voltage across the measurement window prior to decoding, but we found that averaging across time points after decoding tended to yield larger effect sizes (see supplementary Figure S1).

The univariate analyses used data only from the cluster of electrode sites where the component was known to be largest, whereas the SVM analyses used all of the electrode sites, even those at which the component was near zero. We tested the possibility that this might give the univariate analyses an advantage. Specifically, we repeated the SVM analyses using only the electrode sites used for the univariate analyses (listed in Table 1). We found that this approach led to similar or worse results than using all of the electrode sites for decoding (see supplementary Figures S5–S7 and Table S2). We therefore used all of the electrode sites for the decoding analyses reported here.

Table 2. Decoding parameters for all analyses done on each component. The number of crossfolds was determined so that 10-18 trials were included in the averaged ERPs for a given component. The t values shown here represent one-sample t tests against chance for the average decoding accuracy across the measurement window for a given component.

	SVM Parameters								
ERP Component	Avg # of Decoding Trials	Std Deviation	# of Crossfolds	Min # of Trials Per Crossfold	N	df	Decoding t	Decoding p	Time Locking Point
N170 Faces vs. Cars	68.41	6.95	5	13	37	36	44.83	$p < 0.0001$	stimulus onset
N170 Faces vs. Scrambled Faces	67.43	6.87	5	13	37	36	51.02	$p < 0.0001$	stimulus onset
N170 Cars vs. Scrambled Cars	67.16	8.05	5	13	37	36	38.54	$p < 0.0001$	stimulus onset
N170 Scrambled Faces vs. Scrambled Cars	66.65	7.63	5	13	37	36	30.17	$p < 0.0001$	stimulus onset
N400	51.05	5.08	4	10	37	36	45.84	$p < 0.0001$	stimulus onset
N2pc	57.54	9.87	5	11	35	34	6.51	$p < 0.0001$	stimulus onset
LRP	84.15	7.41	5	16	37	36	24.62	$p < 0.0001$	response onset
ERN	41.17	21.12	4	10	36	35	42.10	$p < 0.0001$	response onset
P3	33.62	4.05	3	11	34	33	44.51	$p < 0.0001$	stimulus onset
MMN	185.77	13.94	10	18	39	38	49.94	$p < 0.0001$	stimulus onset

2.7 Crossnobis Analyses

As illustrated in Figure 1C, the Mahalanobis distance is defined as the distance between the centroids of the two classes, scaled by the multidimensional standard deviation of each class. Whereas SVMs rely on good estimates of the values near the decision line to work well, which is

facilitated by averaging, individual data points have less impact on the Mahalanobis distance. However, the Mahalanobis distance benefits from using a large number of data points to obtain robust estimates of the multidimensional standard deviation of a given class, so single-trial data points are typically used instead of averaged ERPs. Because averaging is unnecessary, there is also no need to equate the number of trials in the two conditions being compared, so all the data can be used.

Figure 1c shows a simplified example of crossnobis distance in which ERP data from two electrode sites (e.g., PO8 and Cz) at a particular time point (e.g., 170 ms after stimulus onset) are used to decode whether the subject had been shown a picture of a face or car. The Mahalanobis distance at a given time point is calculated as:

$$Mahalanobis\ Distance = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} * \Sigma^{-1}$$

where x and y correspond to the mean voltages at the two electrodes, the subscripts 1 and 2 correspond to the two conditions (e.g., faces and cars), and Σ corresponds to the multi-dimensional standard deviation across all observations. Note that the centroids and multidimensional standard deviations are computed using exactly the same approach used in linear discriminant analysis.

One shortcoming of the Mahalanobis distance is that it is always positive, so chance is not zero and is difficult to determine. This can be seen in the equation, where the differences between condition means (x_1 - x_2 and y_1 - y_2) are both squared, creating positive values. This shortcoming can be solved by using the crossvalidated Mahalanobis distance (or crossnobis distance; Walther et al., 2016), which involves dividing the data into two evenly random splits, A and B, for each condition. This updated metric for a given time point is then:

$$Crossvalidated\ Mahalanobis\ Distance$$

$$= \sqrt{\left((x_{1A} - x_{2A}) * (x_{1B} - x_{2B})\right) + \left((y_{1A} - y_{2A}) * (y_{1B} - y_{2B})\right)} * \Sigma^{-1}$$

where the equation is the same as used for the Mahalanobis distance, except rather than computing a single squared difference score across conditions at an electrode site (e.g., $(x_1 - x_2)^2$), two independent difference scores are computed across the subsets (A and B) of data across conditions at an electrode site and then multiplied (i.e., $\left((x_{1A} - x_{2A}) * (x_{1B} - x_{2B})\right)$). Because there is no squaring of values[7], the cross-validated distance can be either positive or negative, and the expected chance value is zero (i.e., if the centroids of the two conditions do not differ). Thus, if the crossnobis distance is significantly different from zero, one can conclude that the two conditions are different.

The cross-validation procedure was repeated 100 times with different random splits of the data to provide a more robust result. Each iteration produced a distance value, and we simply averaged these distance values across iterations. Note that, unlike SVM decoding accuracy, the crossnobis distance is not affected by differences in the number of trials in each class. Consequently, we used all trials except for those with behavioral errors or artifacts.

Like the decoding analysis, this analysis was performed separately at each time point for a given component, yielding a waveform of distance values across time points for each participant. As for the SVM decoding, we assessed statistical significance via a one-sample t test against chance (zero) at each time point (excluding the prestimulus period), with an FDR correction for multiple comparisons (Benjamini & Yekutieli, 2001). We also averaged the single-

[7] A consequence of this is that the value inside the square root operation can be either positive or negative. To avoid the use of imaginary numbers, the absolute value is taken prior to computing the square root. The original sign of the value inside the square root operation is then applied to the result.

point crossnobis distances over the recommended measurement window for a given component and compared the result to chance (zero) with a one-sample t test. Negative crossnobis values are not meaningful with the present procedure, so these tests were one-tailed. We also attempted a variant of this analysis in which we averaged the timepoints into a mean amplitude prior to computing the crossnobis distance. However, this approach was consistently worse than our original approach (see supplementary Figure S1). Additionally, we attempted this analysis using only the set of electrodes used for the univariate analysis of a given component, but this did not produce a clear advantage over using all electrode sites (see supplementary Figures S8–S10 and Table S3), so we chose to use all sites in the primary analyses. Finally, we computed the effect size (d_z) of the crossnobis distance:

$$d_z = \left(\overline{X_1} - chance\right) \div \sigma(X_1)$$

where X_1 is equal to the mean of the distance values across participants, chance is 0, and $\sigma(X_1)$ is the standard deviation of the distance values across participants. This is exactly the same d_z formula used for the univariate and SVM analyses. Bootstrapping was used to compute standard errors and 95% confidence intervals for d_z, as detailed in the next section.

Table 3. Crossnobis parameters used for each individual component. Here the t-stats are more comparable to the univariate results. There is no indication of number of trials because this technique, like the univariate approach, allows researchers to use all trials.

Crossnobis Parameters

ERP Component	N	df	Xnobis t	Xnobis p	Time Locking Point
N170 Faces vs. Cars	37	36	9.91	$p < 0.0001$	stimulus onset
N170 Faces vs. Scrambled Faces	37	36	10.19	$p < 0.0001$	stimulus onset
N170 Cars vs. Scrambled Cars	37	36	7.99	$p < 0.0001$	stimulus onset
N170 Scrambled: Faces vs. Cars	37	36	3.46	$p = 0.001$	stimulus onset
N400	37	36	9.82	$p < 0.0001$	stimulus onset
N2pc	35	34	4.14	$p = 0.0002$	stimulus onset
LRP	37	36	14.15	$p < 0.0001$	response onset
ERN	36	35	9.76	$p < 0.0001$	response onset
P3	34	33	8.59	$p < 0.0001$	stimulus onset
MMN	39	38	7.08	$p < 0.0001$	stimulus onset

2.8 Bootstrapping Procedures

We used a standard bootstrapping approach (Efron & Tibshirani, 1994) to estimate the standard error of the effect size values, both at individual time points and when averaged across the measurement window for a given component. For a given score (i.e., a voltage in a univariate difference wave, a cronssnobis value, or a decoding accuracy value), we selected a random sample of N of the N participants (sampling from the set of N participants with replacement) and computed the effect size for this random sample. We iterated this process 10,000 times with

different random samples, obtaining an effect size for each iteration, and the standard error was defined as the standard deviation of these 10,000 effect sizes. For the point-by-point analyses, this was done separately at each time point to obtain a standard error at each time point. For the time-window analyses, the scores were averaged across the time window prior to computing the effect size.

We also used bootstrapping to compare the effect sizes between the three analysis methods (univariate, crossnobis, and SVM). Specifically, we computed the percentage difference in effect size between each pair of analysis methods. The percentage difference was defined as $100 \times (A - B) / B$, where A and B represent the effect sizes for two different methods (e.g., SVM decoding accuracy versus univariate difference voltage). We used bootstrapping to estimate the 95% confidence interval for a given percentage difference. For each of 10,000 bootstrapping iterations, we obtained the percentage difference from a random sample of N of the N participants (sampled with replacement), yielding 10,000 estimates of the percentage difference. In idealized data, the bounds of the 95% confidence interval would be defined as the 2.5 and 97.5 percentiles of this set of estimates, but this can lead to bias in real data, especially if the distribution of values is asymmetrical. We therefore used the 95th percentile bias-corrected and accelerated confidence intervals (Diciccio & Efron, 1996), as implemented in R using the package boot (Canty & Ripley, 2022; Davison & Hinkley, 1997b; R Core Team, 2023).

3. Results

We begin the description of the results with analyses of the N170 experiment, which included four different comparisons: faces versus cars, faces versus scrambled faces, cars versus scrambled cars, and scrambled faces versus scrambled cars. This is followed by an analysis of the N400, N2pc, LRP, and ERN components, each of which involved only two conditions. Those

analyses are followed by analyses of the MMN and P3b components, which were obtained in oddball paradigms and therefore involved comparing conditions with very different numbers of trials.

3.1 N170 Point-by-Point Results

Figure 3 shows the N170 results when analyzed at each individual time point (the point-by-point analysis). The left column shows the traditional univariate analysis (the grand average ERP difference wave at the average of the electrode sites listed in Table 1). The classic comparison of faces and cars (Figure 3A) yielded the expected finding of greater N170 amplitude for faces than for cars. This effect was statistically significant from 100-150 ms (after correction for multiple comparisons). Small but significant differences were also present at other times (indicated by the gray shading in Figure 3). Intact faces and cars yielded a significantly more negative voltage than scrambled faces and cars from approximately 100-600 ms (Figures 3B and 3C). Interestingly, scrambled faces produced a significantly more positive voltage than scrambled cars from 136-200 and 230-292 ms; to our knowledge, this effect has not previously been reported.

The second column of Figure 3 shows point-by-point SVM decoding accuracy, averaged across participants. Decoding accuracy was significantly greater than chance across a broader range of time points compared to the traditional univariate analysis of the difference waves. For example, face versus car decoding was significantly above chance from approximately 70 ms post-stimulus until the end of the epoch, whereas the univariate difference between faces and cars did not reach significance until 100 ms post-stimulus and was significant only during a small set of brief time periods. It should be noted, however, that the univariate ERPs at other electrode sites did differentiate between faces and cars at times when the ERPs from the electrodes shown in Figure 3A did not, so the long-lasting decoding is not incompatible with the

ERP data. Similarly, the univariate difference between scrambled faces and scrambled cars (Figure 3D) was statistically significant in two brief periods, whereas a sustained significant decoding effect was present from 136-312 ms and from 343-363 ms.

The third column of Figure 3 shows the crossnobis distance results, which contained broad periods of statistical significance (93-339 ms) like those obtained for SVM decoding. Note that crossnobis distance is in units of microvolts because it is a distance between the voltage distributions for the conditions being contrasted. However, this metric is scaled by the multivariate standard deviation, so the crossnobis distances are not directly comparable to the univariate ERP amplitudes (even though both are in units of microvolts). Decoding accuracy is expressed in proportion correct, which is even more difficult to relate to the units of the univariate voltage difference and the crossnobis distance. Statistical significance is in the same units for all three analysis methods, but significance is a categorical variable and therefore not a very sensitive metric for comparing the three methods. In addition, one-tailed t test were performed for the decoding measures whereas two-tailed t tests were needed for the univariate voltage measure.

To directly compare the three analysis methods, we converted each of them into the effect size (Cohen's d_z) at each time point. This metric is continuous rather than categorical, and it is directly related to statistical power (Cohen, 1988). The effect sizes are shown in the rightmost column of Figure 3. We took the absolute value of each d_z value for the univariate voltage so that those effect sizes would be positive whether the difference wave contained a positive voltage (e.g., the P3b effect) or a negative voltage (e.g., the N400 effect). The multivariate methods produced positive values for a true effect regardless of the polarity of the effect, so we did not take the absolute value for these methods.

In most cases, the effect size was largest for SVM decoding accuracy, which was more than three times as large as the effect size for the univariate ERPs in some cases (e.g., from 133-172 ms for the faces vs. scrambled faces comparison). The crossnobis effect size tended to be smaller than the SVM decoding effect size but larger than the univariate effect size. The one exception was the comparison of scrambled faces and scrambled cars, for which the effect sizes were similar across the three analysis methods. The general conclusion from these point-by-point analyses is that the multivariate techniques were at least as sensitive as the traditional univariate approach and often considerably more sensitive.

Although univariate analyses sometimes examine point-by-point differences between conditions, as in mass univariate analyses (Groppe et al., 2011; Maris & Oostenveld, 2007), it is more common for univariate analyses to examine the mean voltage across a measurement

window. The next set of analyses compared this time-window univariate approach to decoding and crossnobis values averaged across the same time window.

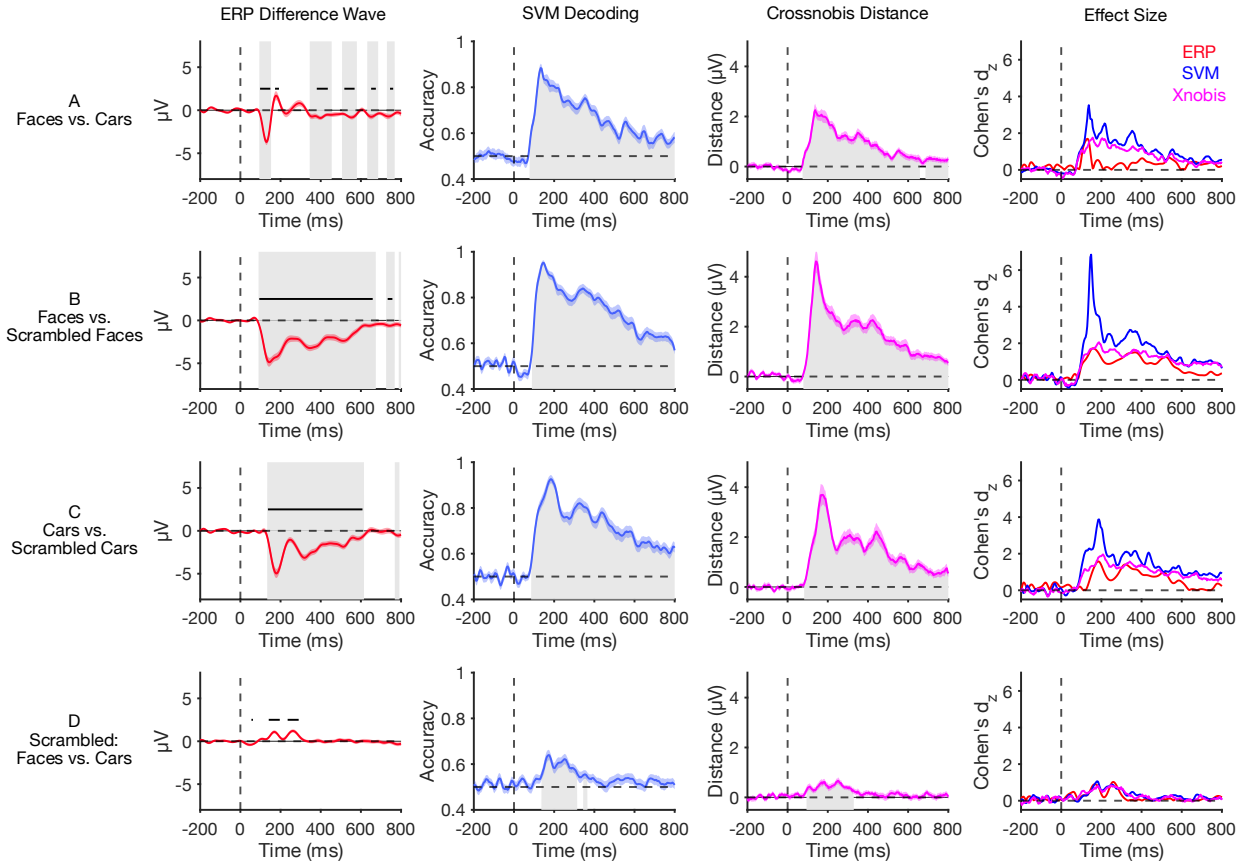


Figure 3. Point-by-point results for the N170 task. The four columns show the mean ERP difference voltages (averaged across the cluster of electrodes listed in Table 1), mean SVM decoding accuracy values, mean crossnobis distances, and effect sizes at each time point. The shaded bands around a given waveform show the standard error of the mean. Gray shading indicates time points at which the value is significantly different from chance with a one-tailed test in the direction of the expected effect after correction for multiple comparisons. The horizontal lines for the ERP difference waves indicate time periods at which the value is significantly different from chance with a two-tailed test (which can reveal effects with an unexpected polarity).

3.2 N170 Time-Window Results

Figure 4 shows the results from the N170 paradigm, averaged across a time window of 110–150 ms (which was the time window recommended for the N170 component; Kappenman

et al., 2021). Each of these effects was significantly different from chance (see Tables 1-3 for individual statistical tests). The effect size for SVM decoding accuracy was 2–3 times as large as the effect size for the univariate voltage difference for the comparisons of faces versus cars (Figure 4A), faces versus scrambled faces (Figure 4B), and cars versus scrambled cars (Figure 4C). The effect size for crossnobis distance was substantially smaller than the effect size for SVM decoding accuracy in these three cases but was as large or larger than the univariate effect size. As was observed in the point-by-point analyses (Figure 3), the effect sizes were similar across the three analysis methods for the comparison of scrambled faces and scrambled cars (Figure 4D), although the effect size was slightly larger for the crossnobis multivariate method than for the univariate method.

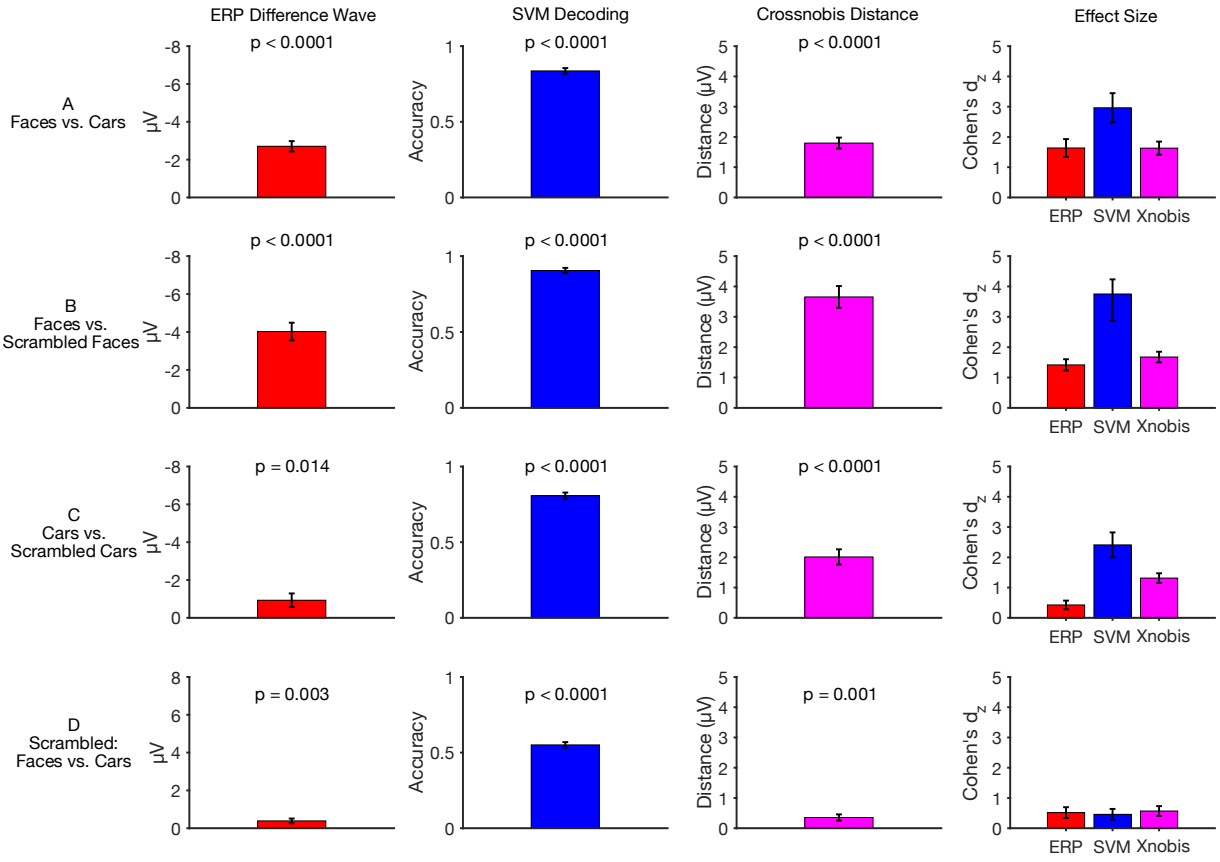


Figure 4. Time-window univariate ERP difference voltage, SVM decoding accuracy, and crossnobis distance, averaged across participants, for the four key comparisons in the N170 task. The time window was 110-150 ms for all three analysis methods. The p values are from one-tailed tests against chance. The right column shows the effect size (Cohen's d_z) for each analysis method. Error bars show ±1 SEM.

3.3 N400, N2pc, LRP, and ERN Results

Figure 5 shows the point-by-point results for the N400, N2pc, LRP, and ERN components, and Figure 6 shows the results averaged over the recommended measurement window for each component.

3.4 N400

In the N400 data (Figures 5A and 6A), the univariate comparison of related versus unrelated words yielded a prototypical negative N400 deflection that became statistically significant at 200 ms, peaked at approximately 400 ms, and remained significant throughout the remainder of the epoch. The SVM decoding and crossnobis distance analyses also sought to distinguish between related and unrelated words and yielded a similar time course. However, the effect sizes shown in the rightmost column of Figure 5A indicate that the effect size was much larger for SVM decoding accuracy than for the univariate voltage difference. By contrast, the effect size was smaller for the crossnobis distance than for the univariate voltage difference. The same pattern of effect sizes was obtained when the data were averaged across the N400 measurement window (300 to 500 ms; see Figure 6A).

3.5 N2pc

In the N2pc paradigm (Figure 5B), the contralateral-minus-ipsilateral difference wave yielded the typical pattern of a negative N2pc that was significant from 175-300 ms, followed by a brief post-N2pc positivity that was significant from 324-355 ms and then a sustained posterior contralateral negativity (SPCN; Kappenman & Luck, 2012; Mazza et al., 2007; Prime &

Jolicoeur, 2010) that was significant from 593-800 ms. For the two multivariate methods, we decoded whether the target was contralateral or ipsilateral to the electrode sites. The midline sites were not used (Fz, FCz, Cz, CPz, Pz, Oz).

For SVM decoding, we first created averages of pairs of trials, one trial with the target in the left visual field (LVF) and one trial with the target in the right visual field (RVF). We averaged the left hemisphere channels for the RVF target trial with the right hemisphere channels for the LVF target trial to create a set of contralateral channels, and we averaged the left hemisphere channels for the LVF target trial with the right hemisphere channels for the RVF target trial to create a set of ipsilateral channels. We then performed the SVM decoding procedure described in Section 2.6, using contralateral and ipsilateral as the classes being decoded.

The crossnobis analysis also used these contralateral and ipsilateral classes, but we needed to ensure that different trials were used for the two splits of the data to avoid artificially minimizing the difference between the two splits. Specifically, the splits were defined by the odd and even trials. In all other respects, the analysis was the same as described in section 2.7. We did this for each color independently to avoid any possible sensory biases effecting the decoding, and then collapsed across the colors.

SVM decoding accuracy became statistically significant at 210 ms, which was slightly later than the univariate effect, and then remained significant until 585 ms. There were also brief periods of significance toward the end of the epoch, from 687-710 ms and from 753-789 ms. The crossnobis distance became significant at 226 ms and remained significant until 523 ms. The effect size (rightmost column of Figure 5B) was larger for the univariate analysis than for the multivariate analyses from 160-265 ms, after which it was larger for the SVM analyses from

approximately 265-585 ms, and then became larger for the univariate analysis again during the SPCN period (from approximately 600 ms through the end of the epoch). The crossnobis analyses was larger than the N2pc analyses from 280-562 ms but not larger than the SVM. When the data were averaged across the N2pc measurement window (200 to 275 ms), the effect size was slightly larger for the univariate voltage than for the two multivariate methods (Figure 6B).

Note that the univariate ERP waveform contained a sequence of a negative N2pc followed by a post-N2pc positivity, which are thought to reflect two distinct neurocognitive processes (Gaspelin et al., 2023). However, there was a single period spanning these two components in which the SVM decoding accuracy and crossnobis distance was significantly greater from chance. This illustrates a potential downside of using multivariate pattern analysis methods to analyze ERP data, namely the loss of information about different polarities that might indicate different ERP components.

3.6 LRP

In the flankers paradigm, the response-locked LRP analysis (Figure 5C) yielded a negative voltage in the contralateral-minus-ipsilateral difference wave that began approximately 128 ms prior to the response and lasted until 113 ms after the response, at which point the voltage became positive. There were also isolated periods of small but significant contralateral positivity prior to the negativity.

The multivariate analyses were performed using the same approach used for the N2pc, except that the data were organized according to whether the response was contralateral versus ipsilateral relative to the electrode sites. SVM decoding accuracy was slightly but significantly greater than chance from 324 to 226 ms prior to the response. This was followed by a stronger period of decoding that was significant from 160 ms prior to response until the end of the epoch.

A sustained period of significant crossnobis distance began 180 ms prior to the response and lasted through the end of the epoch.

The effect size was similar for SVM decoding accuracy and the univariate voltage during the early part of the LRP (approximately 150-62 ms prior to the response), and then became greater for SVM decoding accuracy until the end of the epoch. The crossnobis effect size was substantially smaller than the univariate effect size prior to the response and then became similar after the response. When the data were averaged across the LRP measurement window (-100 to 0 ms), the effect size was slightly larger for the SVM decoding accuracy than for the univariate voltage and substantially lower for the crossnobis distance (Figure 6C).

3.7 ERN

Figure 5D shows the ERN analyses from the flankers paradigm, which examined the difference in voltage between error trials and correct trials. As usual, this analysis yielded a more negative voltage on error trials than on correct trials, which became significant 35 ms before the response and remained significant until 110 ms after the response. The ERN was followed by an error positivity (Pe) that was significant from 132 ms until the end of the epoch. There were also three brief periods of significance prior to the ERN, perhaps reflecting differences in perceptual and cognitive processes that lead to a correct response versus an incorrect response.

The SVM and crossnobis approaches attempted to decode whether the response was correct or incorrect. SVM decoding accuracy and the crossnobis distance became significantly different from chance during this early period prior to the ERN (beginning at 266 ms pre-response for the SVM and at203 ms pre-response for the crossnobis analysis), and they remained significant until the end of the epoch. Note that the univariate analysis made it possible to see separate ERN and Pe components with different polarities, whereas the multivariate analysis did

not make this distinction. Prior to the response, the effect size was approximately equal for the two multivariate methods and somewhat smaller for the univariate effect size. After the response, the effect size was greater for the SVM decoding accuracy than for the other two methods. When the data were averaged across the ERN measurement window (0 to 100 ms), the effect size was substantially larger for the SVM decoding accuracy than for the univariate voltage and the crossnobis distance (Figure 6D).

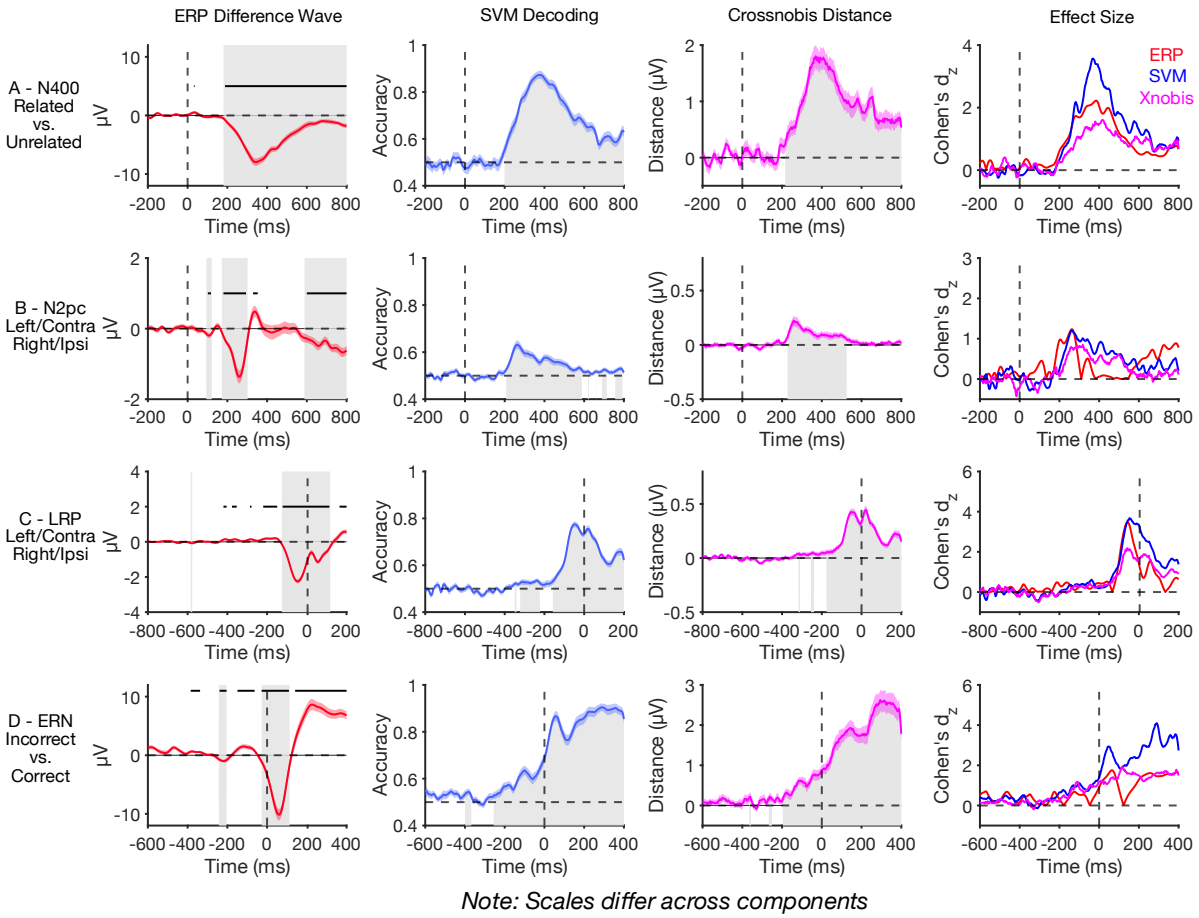


Figure 5. Point-by-point results for the N400, N2pc, LRP, and ERN components. The four columns show the mean ERP difference voltages (averaged across the cluster of electrodes listed in Table 1), mean decoding accuracy values, mean crossnobis distances, and effect sizes at each time point. The shaded bands around a given waveform show the standard error of the mean. Gray shading indicates time points at which the value is significantly different from chance with a one-tailed test in the direction of the expected effect after correction for multiple comparisons. The horizontal lines for the ERP difference waves indicate time periods at which the value is significantly different from chance with a two-tailed test (which can reveal effects with an unexpected polarity).

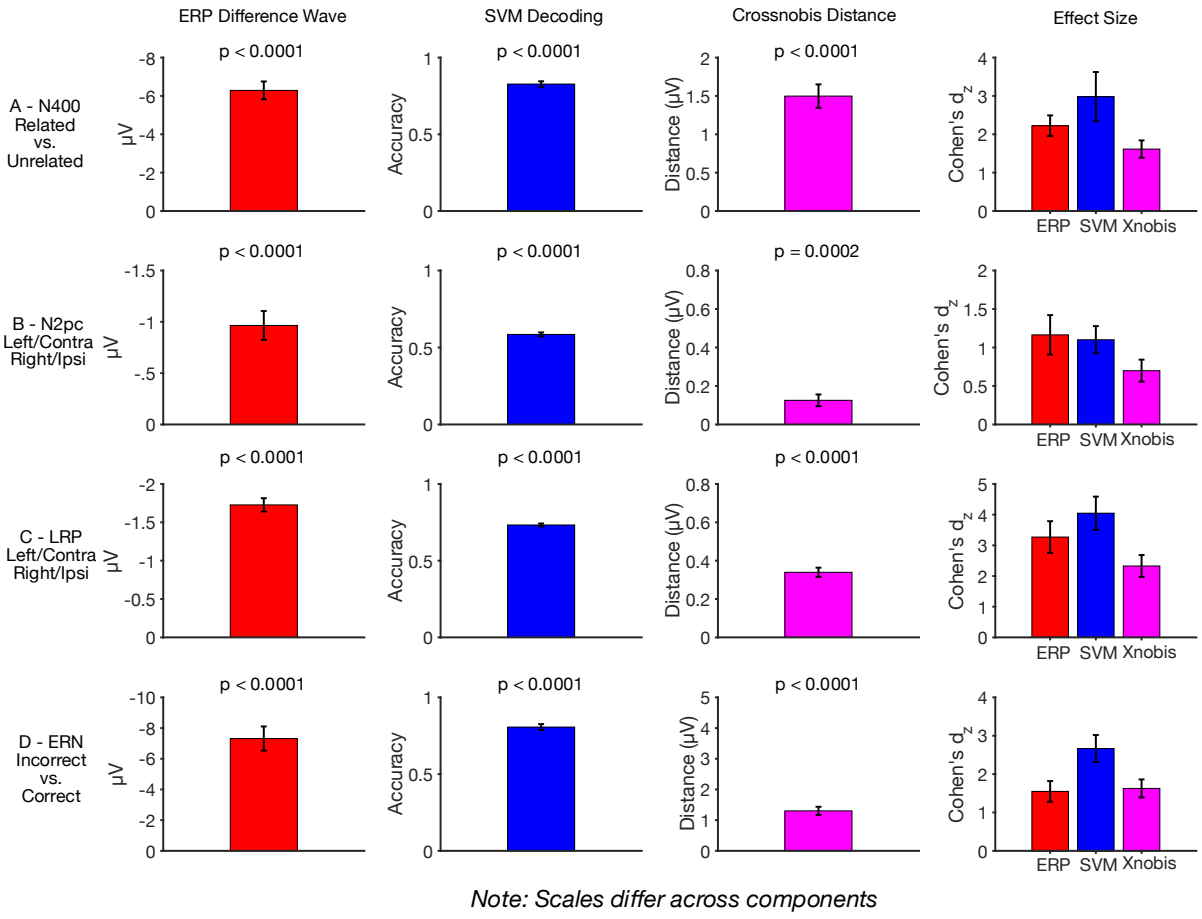


Figure 6. Time-window univariate ERP difference voltage, SVM decoding accuracy, and crossnobis distance, averaged across participants, for the N400, N2pc, LRP, and ERN components. The measurement window was 300 to 500 ms relative to the stimulus for the N400, 200 to 275 ms relative to the stimulus for the N2pc, -125 to 0 ms relative to the response for the LRP, and 0 to 100 ms relative to the response for the ERN. The p values are from one-tailed tests against chance. The right column shows the effect size (Cohen's d_z) for each analysis method. Error bars show ±1 SEM.

3.8 P3b and MMN Results

We chose to describe the results for the P3b and MMN components in the same section because both were assessed using oddball paradigms that are explicitly designed to include one condition with many fewer trials than the other condition. The SVM and crossnobis methods were used to decode whether the stimulus came from the rare category or the frequent category.

Given that our SVM decoding approach requires equating the number of trials across categories by subsampling from the more frequent category, these paradigms may be a special case.

3.9 P3b

In the P3b experiment, the univariate data yielded a prototypical P3b effect, consisting of a more positive voltage for the rare targets than for the frequent standards, peaking at approximately 380 ms (see Figure 7A). This effect was significant from 260–660 ms. It was preceded by a much smaller positivity— a frontally-maximal P2 wave—that was significant from 200-226 ms. There was also a brief period of significance from 42-70 ms that is presumably spurious due to its implausibly early onset.

The SVM decoding accuracy and crossnobis distance waveforms significantly exceeded chance beginning just before 200 ms, during the time period of the frontal P2 effect, and remained significant through the end of the epoch. During the P2 period, the effect size was much larger for the two multivariate methods than for the univariate voltage, although this may simply reflect the fact that the electrode cluster for the univariate analysis was optimized for the P3b rather than for the more frontally distributed P2. From approximately 300 ms onward, the effect size was similar across methods, but with a small advantage for SVM decoding accuracy near the peak of the P3 wave. When the data were collapsed across the P3b measurement window (300 to 600 ms), the effect size was largest for SVM decoding accuracy, intermediate for the univariate voltage, and smallest for the crossnobis distance (see Figure 8A).

3.10 MMN

In the MMN experiment, the univariate data yielded a prototypical MMN effect, consisting of a more negative voltage for deviants than for standards, peaking at approximately

200 ms (see Figure 7B). This effect was significant from 100–240 ms. There was also a significant negativity from 340–546 ms.

The SVM decoding accuracy waveform significantly exceeded chance from 117–406 ms, and the crossnobis distance waveform showed sustained significance from 117–386 ms. The peak effect size was largest for the univariate voltage, slightly smaller for SVM decoding accuracy, and even smaller for crossnobis distance. This same pattern of effect sizes was observed when the data were collapsed across the MMN measurement window (125 to 225 ms; see Figure 8B).

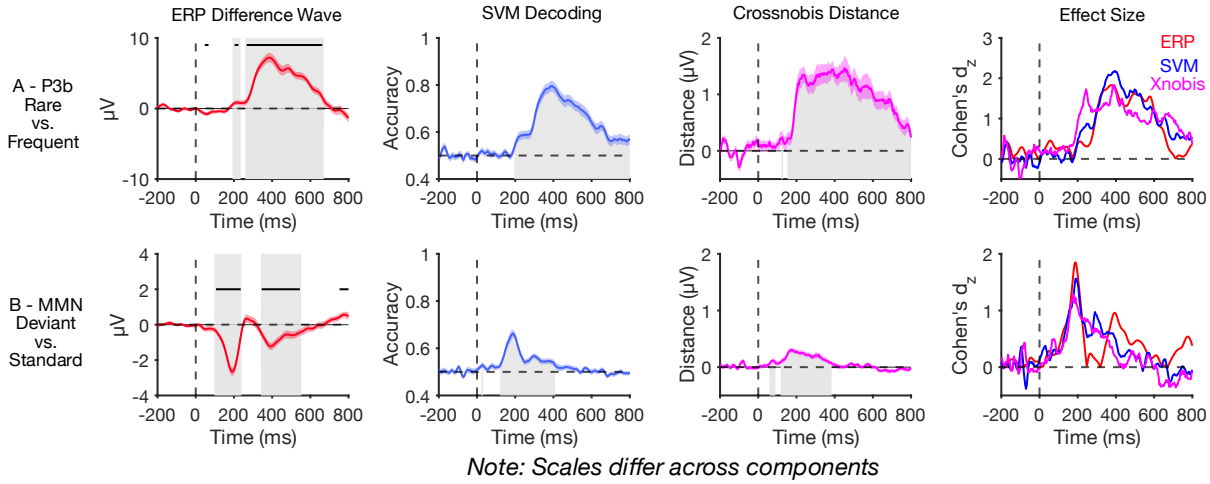


Figure 7. Point-by-point results for the P3b and mismatch negativity (MMN) components. The four columns show the mean ERP difference voltages (averaged across the cluster of electrodes listed in Table 1), mean decoding accuracy values, mean crossnobis distances, and effect sizes at each time point. The shaded bands around the a given waveform show the standard error of the mean. Gray shading indicates time points at which the value is significantly different from chance with a one-tailed test in the direction of the expected effect after correction for multiple comparisons. The horizontal lines for the ERP difference waves indicate time periods at which the value is significantly different from chance with a two-tailed test (which can reveal effects with an unexpected polarity).

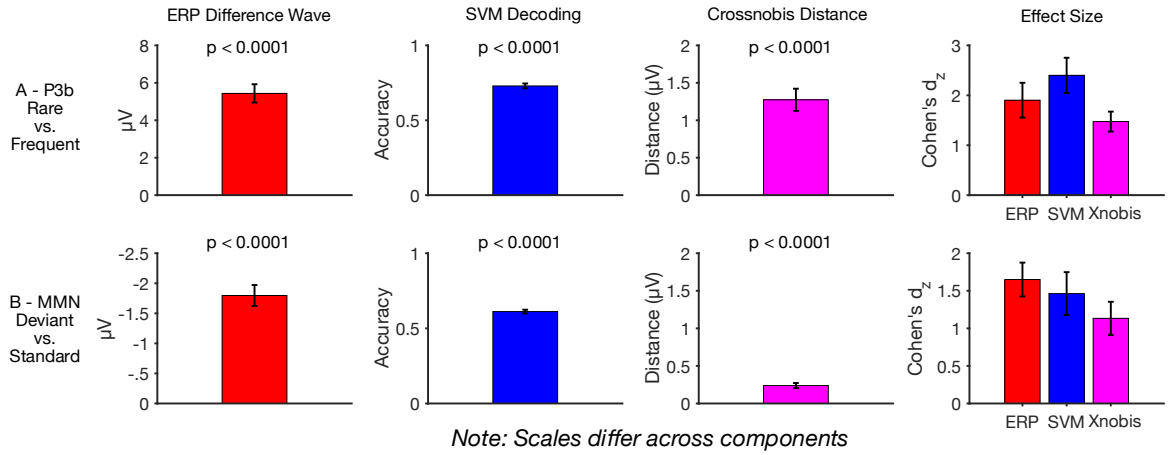


Figure 8. Time-window univariate ERP difference voltage, decoding accuracy, and crossnobis distance, averaged across participants, for the P3b and MMN components. The measurement window was 300–600 ms for the P3b component and 125–225 ms for the MMN. The p values are from one-tailed tests against chance. The right column shows the effect size (Cohen's d_z) at each time window for each analysis method. Error bars show ±1 SEM.

3.11 Summary of Effect Sizes and Effect Size Contrasts

Figure 9 summarizes the time-window effect sizes for all the analyses, using the same scale for all components to facilitate comparison across components. The effect size was much larger for the SVM decoding accuracy than for the other approaches in several cases and was never much smaller than the univariate effect size.

To provide a direct comparison of the effect sizes, we computed the percentage difference between each pair of analysis methods. The percentage difference was defined as 100 × (A – B) / B, where A and B represent the effect sizes for two different methods (e.g., SVM decoding accuracy versus univariate difference voltage). We then obtained the 95% confidence intervals of these percentage differences. SVM decoding accuracy yielded the largest effect size in seven of the ten comparisons, figure 10. In these seven cases, the effect size was 14–467% greater for SVM decoding accuracy than for the univariate voltage (with a confidence interval that excluded zero for P3b, ERN, N170 faces versus cars, N170 faces versus scrambled faces,

and N170 cars versus scrambled cars). In the other three comparisons, the effect size was only 5%–11% smaller for SVM decoding accuracy than for the univariate voltage, and none of these confidence intervals excluded zero. Thus, SVM decoding accuracy was often much better than the univariate voltage and never much worse, figure 10.

The crossnobis distance metric yielded larger effect sizes than the univariate voltage in 4 out of 10 comparisons, for which the benefit was 5–209%. In the remaining 6 comparisons, the effect size was 0.4–40% smaller for the crossnobis distance than for the univariate voltage. Moreover, the SVM effect size was greater than the crossnobis effect size in 9 out of 10 comparisons, with a 36–124% larger effect size for SVM than for crossnobis in these 9 cases. The scrambled faces vs scrambled cars comparison was the only case in which crossnobis yielded a larger effect size than SVM (a 19% advantage, with a confidence interval that included zero), figure 10.

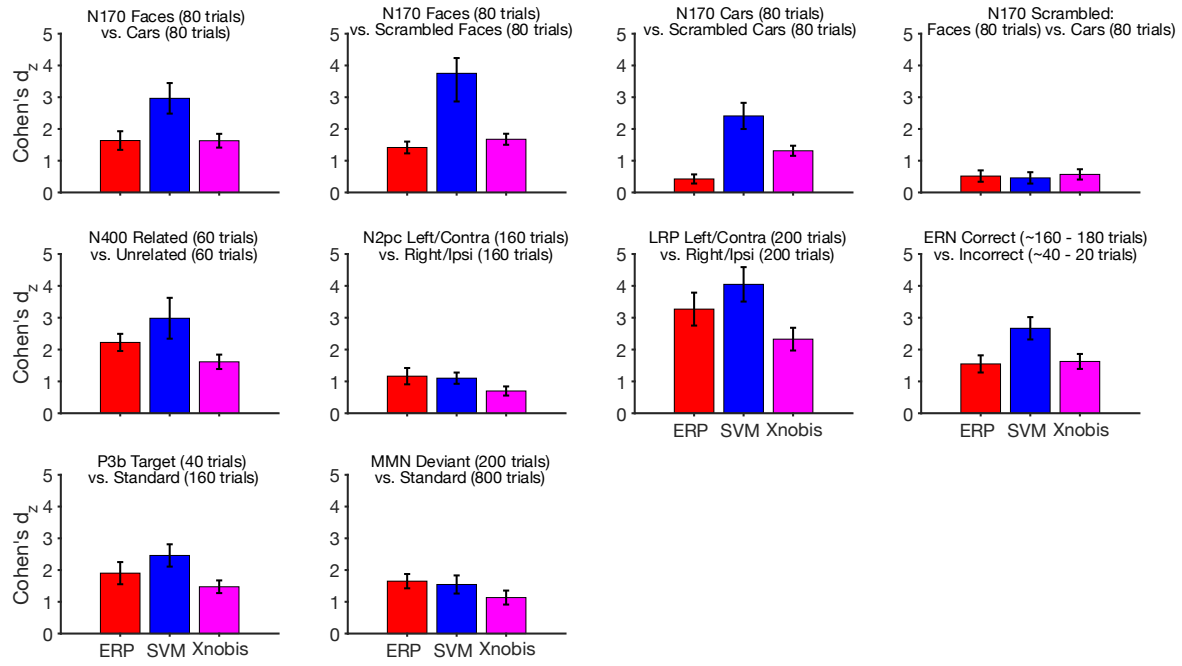
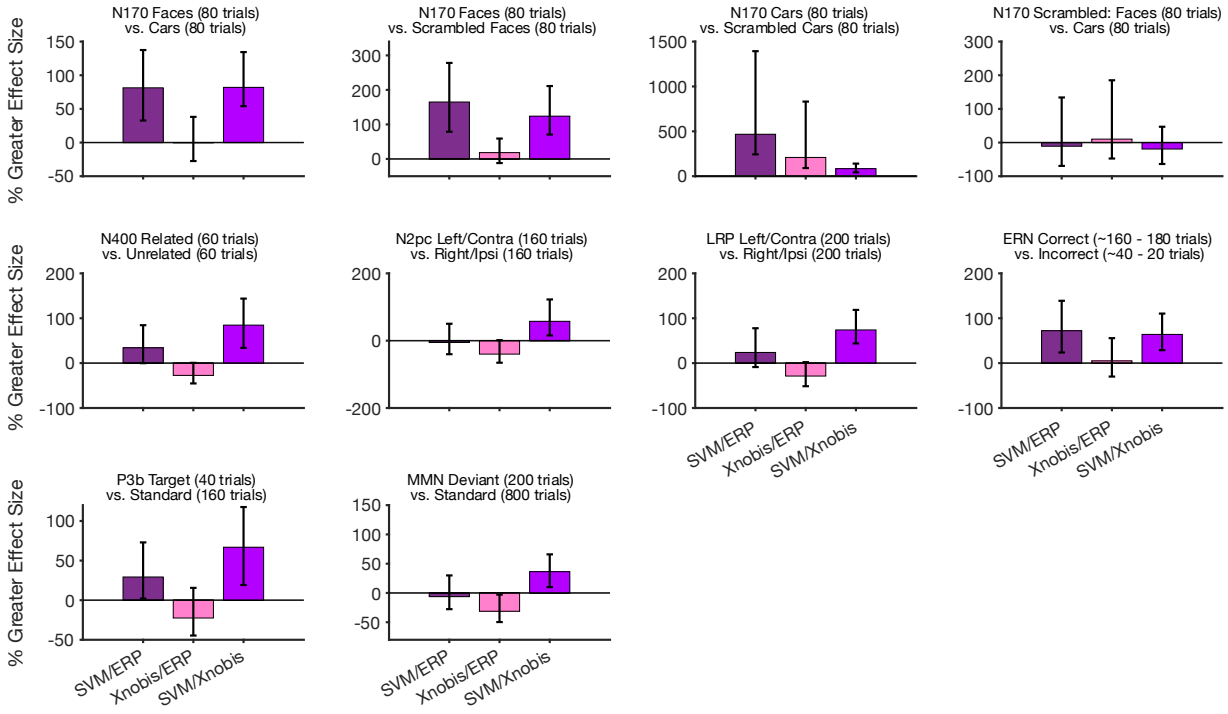


Figure 9. Effect size (Cohen's d_z) for each analysis method for all the time-window comparisons, drawn on the same scale for all components. The error bars show ±1 standard error.



Note: Scales differ across components

Figure 10. Percentage differences in effect sizes between each pair of analysis methods for all time-window comparisons. The bars show 95% confidence intervals.

3.12 Mean Amplitude × Multivariate Time Window Regressions

The univariate and multivariate analysis methods were applied to the same data, and they assessed the same or similar contrasts between experimental conditions, so it seems plausible that participants who have a large univariate difference in voltage between conditions will also exhibit high SVM decoding accuracy and large crossnobis distances. However, the SVM and crossnobis methods also take into account the variability across trials within each participant, so a participant with a large difference in mean voltage between two conditions but high trial-to-trial variability might have low decoding accuracy and a small crossnobis distance. Thus, it is an empirical question whether the size of the univariate voltage difference will be associated with the SVM decoding accuracy or the crossnobis distance. We therefore used simple linear

regressions to assess the strength of these associations. The polarities of the plots and the r values were reversed when necessary, so that a larger univariate effect was always coded as a more positive value.

As shown in Figure 11A, the univariate voltage difference between conditions was significantly correlated with the SVM decoding accuracy across all ten comparisons. The magnitude of the correlations varied widely from largely medium to strong correlations (from r = 0.344 to 0.791), but the correlation was above 0.4 for all components except the N170 faces vs scrambled faces comparison which is where the univariate effect size was quite small.

Figure 11B shows the correlations between the univariate voltage differences and the crossnobis distances. These correlations were much smaller than those for the SVM decoding accuracy, and they failed to reach significance in six of the ten cases (N170 faces vs. scrambled faces; N170 scrambled faces vs. scrambled cars; ERN, P3b, and MMN).

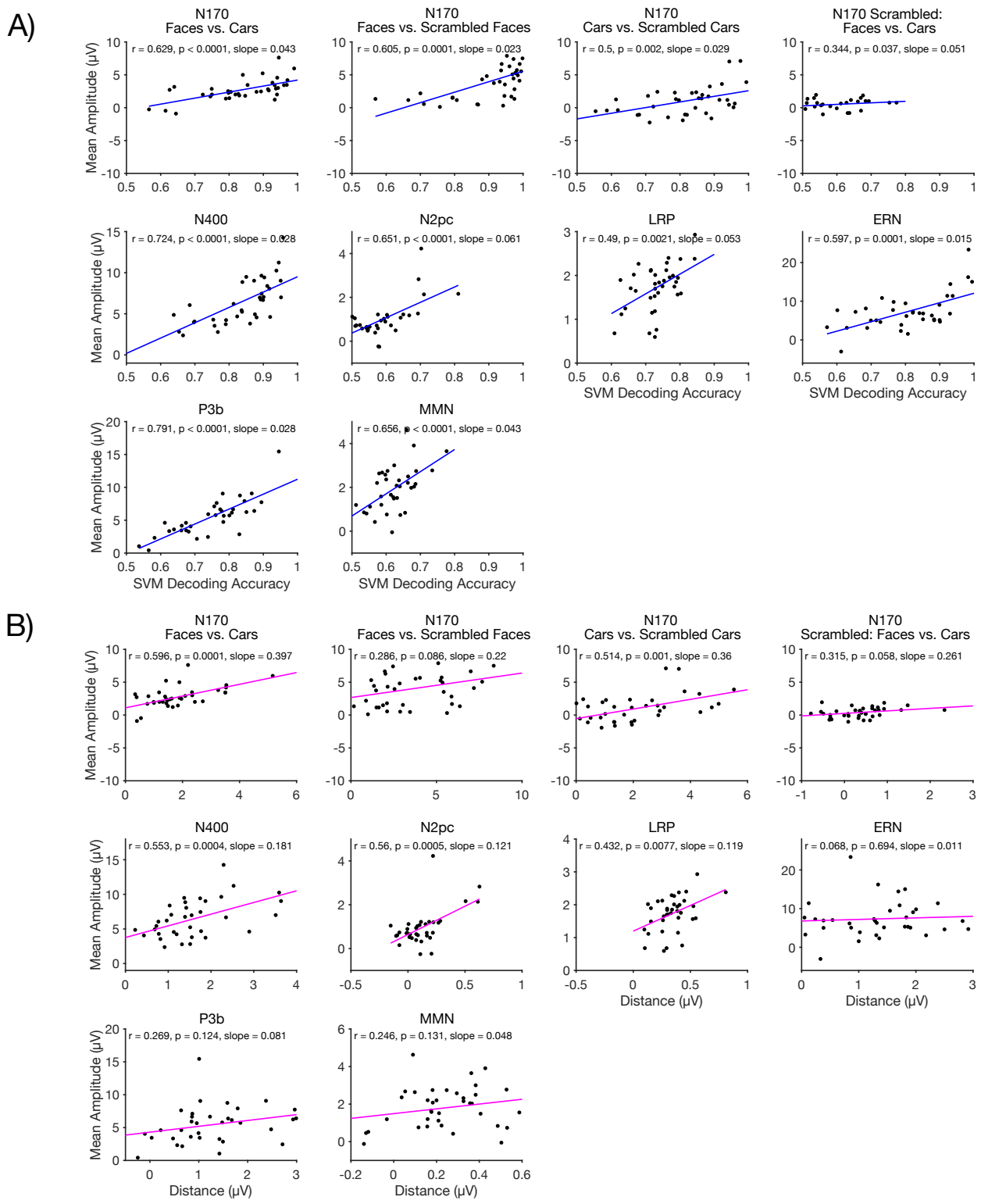


Figure 11. Difference wave mean amplitude as a function of SVM decoding accuracy (A) or crossnobis distance (B). Note that the polarities of the plots and r values were reversed when necessary, so that a larger univariate effect was always coded as a more positive value. The Pearson r correlation coefficient, corresponding p value, and slope of the linear relationship are shown for each comparison.

4. Discussion

In this study, we compared the traditional univariate approach to assessing differences in amplitude between experimental conditions to newer multivariate pattern analysis methods that apply machine learning to the entire topographic pattern of voltage across the scalp. We examined ten different comparisons across seven different ERP components so that we could assess the generality of any differences. When the data were averaged across the measurement window for the component of interest, we found that SVM decoding accuracy produced substantially larger effect sizes than traditional univariate methods in seven of the ten comparisons, with a slightly smaller effect size for SVM decoding accuracy in the other three cases. Similar results were obtained in point-by-point analyses, although there was some variation in the relative effect sizes across time points. In almost every case, effect sizes were lower for the crossnobis distance than for SVM decoding accuracy. These results suggest that ERP researchers could obtain larger effect sizes, and therefore greater statistical power, by using SVM decoding accuracy as their primary dependent variable instead of the univariate voltage, at least for the types of participants and paradigms examined in the present study. Moreover, SVM decoding is now very straightforward to apply in ERPLAB Toolbox (Lopez-Calderon & Luck, 2014), using either scripting or a point-and-click graphical user interface.

Larger effect sizes and greater statistical power would be a substantial boon to ERP research. A larger effect size can decrease the Type II error rate and thereby increase the proportion of significant effects that are actually true effects (Ioannidis, 2005). Alternatively, researchers could achieve the same level of statistical power as before with fewer participants, which would increase the number of studies that are feasible given limited resources. However, there are some important limitations to the present findings and to the use of SVM decoding in

place of traditional univariate analyses, so researchers should carefully consider whether to replace conventional analyses with SVM decoding. We will focus the remainder of our discussion on these limitations.

Although the present study used many different comparisons and ERP components to enhance the generality of our conclusions, the benefits of SVM decoding we observed may not generalize to all ERP research. The ERP CORE data used for the present analyses were recorded using active electrodes with a stable gel-based interface between the electrodes, and the recordings were obtained inside an electrically shielded chamber (which minimizes line noise) with excellent temperature control (which minimizes skin potentials). Moreover, the data were obtained from highly compliant neurotypical young adults. Because SVM decoding accuracy and the crossnobis distance are both decreased when trial-to-trial variability in the EEG is greater, they may yield smaller effect sizes than univariate voltage measures in datasets with greater noise. However, univariate effect sizes are also decreased when the trial-to-trial variability increases. It would therefore be useful for future research to conduct similar comparisons between univariate and multivariate methods in data obtained in noisier recording setups and in data from participant populations who exhibit greater trial-to-trial EEG variability.

A second limit to generality is that we cannot be certain that SVM decoding accuracy would lead to larger effect sizes for all ERP paradigms. Indeed, we found that it led to comparable or slightly smaller effect sizes than the univariate voltage in a few cases. It is not clear why. The fact that SVM decoding was not very helpful for the N2pc and LRP components may reflect the fact that the traditional univariate approach already uses a difference in scalp distribution (contralateral versus ipsilateral) to quantify the amplitude of the component. However, this does not explain why SVM decoding had no advantage over the univariate voltage for the MMN or

for the comparison of scrambled faces and scrambled cars in the N170 experiment. These variations in the relative effectiveness of SVM decoding may reflect some as-yet-to-be-determined variable or they may reflect random variation. It will be important for future research to determine whether there are specific conditions under which SVM decoding is advantageous. However, SVM decoding was never substantially worse than univariate voltage analyses in the present data, and it was often substantially better, so researchers who are analyzing similar data should seriously consider using SVM decoding.

A key difference between traditional univariate analyses and MVPA analyses is that univariate analyses often focus on the subset of electrodes where the effect is largest, whereas MVPA analyses typically use all electrodes. Focusing on electrodes with large effects might be expected to confer an advantage, but we found that including all the sites in the MVPA analyses led to equivalent or even larger effect sizes than including only the sites from the univariate analyses (see supplementary Figures S5 – S10).

Another increasingly common univariate approach is the mass univariate approach, in which conditions are statistically compared separately at each electrode site, and then an intelligent correction for multiple comparisons is applied (Maris & Oostenveld, 2007). There is no straightforward way to assess the effect size with this approach, which makes it difficult to compare with MVPA approaches. However, the statistical power of a mass univariate analysis is typically lower than the power of the univariate approach used in the present study, because confining a univariate analysis to electrode sites that are known a priori to contain the largest effects increases the power of the analysis relative to a mass univariate analysis (Groppe et al., 2011). Thus, we would expect that SVM-based decoding analyses would have even greater

advantages over mass univariate analyses than over the simple univariate analyses examined here.

A related issue is that univariate analyses provide scalp distribution information, whereas SVM-based decoding does not. For example, studies using univariate approaches often include scalp maps of the effects (including all electrodes, not just those used in the analyses). Moreover, the mass univariate approach can indicate which sites exhibit statistically significant effects. This information is not provided by SVM-based decoding. Although it is possible to look at the weight for each electrode site in an SVM analysis, Haufe et al. (2014) showed that this can be misleading. They proposed using a transformation of the weights, but this transformation applies only to linear discriminant analysis, not to SVM-based decoding. In addition, MVPA approaches are applied separately to each participant, so there is no group-level scalp map. However, this single-participant approach is also an advantage, because it is not hampered by differences across participants in the scalp distribution of an effect, which may be quite large due to idiosyncrasies in biophysical factors (Hajizadeh et al., 2021). This may be a key reason why SVM-based decoding often led to larger effect sizes than univariate methods in the present study.

The fact that SVM decoding requires equating the number of trials between the conditions being compared is a potential downside and may explain the relatively poor performance of SVM decoding for the MMN (see Luck, 2023 for an extended discussion). However, we obtained larger effect sizes for SVM decoding than for the univariate voltage for the P3b and ERN components, which also required discarding large numbers of trials in one of the two conditions being compared. It should be noted that care would be needed in applying SVM decoding in studies comparing groups of participants that differ in the number of trials available for averaging (e.g., due to differences in the number of trials rejected because of artifacts). In

such studies, it would be necessary to match the number of trials per groups by subsampling from the available trials. This could be accomplished by pairing participants across groups and using subsampling to equate the number of trials in each pair.

Increased trial-to-trial variability does not impact the expected value of traditional mean amplitude measures (Luck, 2014). That is, although the measured mean amplitude may be less representative of the true value in noisier data, it is not consistently larger or consistently smaller in noisier data than in cleaner data. However, the expected value for SVM decoding accuracy becomes progressively smaller as the trial-to-trial variability increases. As a result, it can be problematic to compare decoding accuracy across groups or experimental conditions that differ in the amount of trial-to-trial variability. For example, it may be problematic to compare decoding accuracy between a patient group and a control group if the patient group has more movement artifacts than the control group. A method has been developed to quantify the specific type of trial-to-trial variability that impacts SVM decoding accuracy (Bae et al., 2023), and a method of this nature should be used whenever there is reason to believe that groups or conditions might differ in their variability. However, this is not a problem if decoding accuracy is actually greater in the group with greater variability than the group with less variability. It should also be noted that this issue is not unique to decoding accuracy: The expected value of traditional peak amplitude measures increases progressively as the noise level increases, making it problematic to compare groups with different noise levels (unless the group with a higher noise level is found to have lower peak amplitudes than the group with a lower noise level).

Another important limitation of decoding is the loss of polarity information. In the case of the N2pc component, for example, a contralateral positivity often follows the N2pc and is thought to reflect a different cognitive process than the N2pc (Gaspelin et al., 2023). However,

the SVM and crossnobis analyses are blind to polarity and simply indicate that the two conditions are discriminable across a set of time points.

On the other hand, it can be very difficult to isolate a specific ERP component and to know whether the same component is being observed across different experimental paradigms and participant populations (Kappenman & Luck, 2012b). Decoding approaches avoid these thorny issues, instead asking whether two experimental conditions yield discriminably different patterns of neural activity, how different they are, and the latencies at which they are discriminable. Indeed, the interpretation of decoding accuracy is subtly but importantly different than the traditional interpretation of univariate voltage differences. Univariate ERP analyses are traditionally used to quantify the extent to which a specific ERP component differs across conditions, with the implicit assumption that the component has a relatively constant scalp distribution across participants (and can therefore be measured at the same electrode or cluster of electrodes in all participants). Decoding analyses instead quantify the amount of information about condition that is present in the recorded signal in each individual participant, with no assumption that the underlying brain activity is similar across participants. These are both valid things to quantify, and which one is more useful will depend on the scientific question being asked by a given study. Researchers must therefore think carefully about whether their scientific question requires isolating a specific component or whether it is instead related to the information content of the neural signal.

Data and Code Availability

All raw and preprocessed univariate data used for this project are a part of the ERP CORE resource, which is available at https://doi.org/10.18115/D5JW4R. The data and code used to implement all decoding analyses that is compatible with ERPLAB's GUI is available here: doi.org/10.17605/OSF.IO/5WV93.

Author Contributions

Carlos Daniel Carrasco: conceptualization, analyses, writing; Brett Bahle: development of crossnobis method, crossnobis description writing, Aaron Matthew Simmons: software development; Steven J. Luck: conceptualization, writing, development of methods, funding acquisition. Editing/Review was performed by all authors prior to submission of the final version of the manuscript.

Acknowledgements

We thank John Nadra for always providing feedback and support throughout the duration of this project and many others. We also thank the members of the Luck Lab for their valuable contributions to community and thoughtful insights, including but not limited to: David Garrett, Guanghui Zhang, Lara Krisst, Kurt Winsler, and John Kiat.

Funding

This study was supported by grants R01MH087450 and R01EY033329 from the National Institute of Mental Health.

References

Ashton, K., Zinszer, B. D., Cichy, R. M., Nelson, C. A., Aslin, R. N., & Bayet, L. (2022). Time-resolved multivariate pattern analysis of infant EEG data: A practical tutorial. Developmental Cognitive Neuroscience, 54. https://doi.org/10.1016/j.dcn.2022.101094

Bae, G. Y., Leonard, C. J., Hahn, B., Gold, J. M., & Luck, S. J. (2023). Assessing the information content of ERP signals in schizophrenia using multivariate decoding methods. NeuroImage: Clinical, 25. https://doi.org/10.1016/j.nicl.2020.102179

Bae, G. Y., & Luck, S. J. (2019). Decoding motion direction using the topography of sustained ERPs and alpha oscillations. NeuroImage, 184, 242–255. https://doi.org/10.1016/j.neuroimage.2018.09.029

Bae, G. Y. (2021). The Time Course of Face Representations during Perception and Working Memory Maintenance. Cerebral Cortex Communications, 2(1), 1–12. https://doi.org/10.1093/texcom/tgaa093

Bae, G. Y., & Luck, S. J. (2018). Dissociable Decoding of Spatial Attention and Working Memory from EEG Oscillations and Sustained Potentials. The Journal of Neuroscience, 38(2), 2860–17. https://doi.org/10.1523/JNEUROSCI.2860-17.2017

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics, 29(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Borrell, V. (2018). How cells fold the cerebral cortex. Journal of Neuroscience, 38(4), 776–783. https://doi.org/10.1523/JNEUROSCI.1106-17.2017

Canty, A., & Ripley, B. D. (2022). boot: Bootstrap R (S-Plus) Functions.

Chan, A. M., Halgren, E., Marinkovic, K., & Cash, S. S. (2011). Decoding word and category-specific spatiotemporal representations from MEG and EEG. NeuroImage, 54(4), 3028–3039. https://doi.org/10.1016/j.neuroimage.2010.10.073

Cohen J. (1988). Statistical Power Analysis for the Behavioral Sciences Second Edition.

Collins, A. G. E., & Frank, M. J. (2018). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. Proceedings of the National Academy of Sciences of the United States of America, 115(10), 2502–2507. https://doi.org/10.1073/pnas.1720963115

Daly, I. (2023). Neural decoding of music from the EEG. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-022-27361-x

Davison, A. C., & Hinkley, D. V. (1997). Bootstrap Methods and Their Applications. Cambridge University Press. http://statwww.epfl.ch/davison/BMA/

de Vries, I. E. J., van Driel, J., & Olivers, C. N. L. (2019). Decoding the status of working memory representations in preparation of visual selection. NeuroImage, 191, 549–559. https://doi.org/10.1016/j.neuroimage.2019.02.069

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. In Journal of Neuroscience Methods (Vol. 134). http://www.sccn.ucsd.edu/eeglab/

Diciccio, T. J., & Efron, B. (1996). Bootstrap Confidence Intervals. In Statistical Science (Vol. 11, Issue 3).

Dietterich, T. G., & Bakiri, G. (1995). Solving Multiclass Learning Problems via Error-Correcting Output Codes. In Journal of Artiicial Intelligence Research (Vol. 2).

Dobs, K., Isik, L., Pantazis, D., & Kanwisher, N. (2019). How face perception unfolds over time. Nature Communications, 10(1). https://doi.org/10.1038/s41467-019-09239-1

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

Eimer, M., & Coles, M. G. H. (2003). The lateralized readiness potential. The Bereitschaftspotential: Movement-Related Cortical Potentials, 229–248.

Ester, E., & Weese, R. (2023). Temporally Dissociable Mechanisms of Spatial, Feature, and Motor Selection during Working Memory-guided Behavior. Journal of Cognitive Neuroscience, 1–14. https://doi.org/10.1162/jocn_a_02061

Fernández, V., Llinares-Benadero, C., & Borrell, V. (2016). Cerebral cortex expansion and folding: what have we learned? The EMBO Journal, 35(10), 1021–1044. https://doi.org/10.15252/embj.201593701

Gaspelin, N., Lamy, D., Egeth, H. E., Liesefeld, H. R., Kerzel, D., Mandal, A., Müller, M. M., Schall, J. D., Schubö, A., Slagter, H. A., Stilwell, B. T., & van Moorselaar, D. (2023). The Distractor Positivity Component and the Inhibition of Distracting Stimuli. Journal of Cognitive Neuroscience, 1–23. https://doi.org/10.1162/jocn_a_02051

Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (2012). The error-related negativity (ERN/Ne). Oxford Handbook of Event-Related Potential Components, 231–291.

Gholami, R., & Fakhari, N. (2017). Support Vector Machine: Principles, Parameters, and Applications. In Handbook of Neural Computation (pp. 515–535). Elsevier Inc. https://doi.org/10.1016/B978-0-12-811318-9.00027-2

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. Journal of Cognitive Neuroscience, 29(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. In Psychophysiology (Vol. 48, Issue 12, pp. 1711–1725). Blackwell Publishing Inc. https://doi.org/10.1111/j.1469-8986.2011.01273.x

Gurariy, G., Mruczek, R. E. B., Snow, J. C., & Caplovitz, G. P. (2022). Using High-Density Electroencephalography to Explore Spatiotemporal Representations of Object Categories in Visual Cortex. Journal of Cognitive Neuroscience, 34(6), 967–987. https://doi.org/10.1162/jocn_a_01845

Hajizadeh, A., Matysiak, A., Brechmann, A., König, R., & May, P. J. C. (2021). Why do humans have unique auditory event-related fields? Evidence from computational modeling and MEG experiments. Psychophysiology, 58(4). https://doi.org/10.1111/psyp.13769

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage, 87, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Hubbard, J., Kikumoto, A., & Mayr, U. (2019). EEG Decoding Reveals the Strength and Temporal Dynamics of Goal-Relevant Representations. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-019-45333-6

Ioannidis, J. P. A. (2005). Why most published research findings are false. In Why most published research findings are false. (Vol. e124, pp. 2–8). Springer International Publishing. https://doi.org/https://doi.org/10.1371/journal.pmed.0020124

Kappenman, E. S., Farrens, J. L., Zhang, W., Stewart, A. X., & Luck, S. J. (2021). ERP CORE: An open resource for human event-related potential research. NeuroImage, 225(May 2020), 117465. https://doi.org/10.1016/j.neuroimage.2020.117465

Kappenman, E. S., & Luck, S. J. (2012a). Electrophysiological correlates of the focusing of attention within complex visual scenes: N2pc and related ERP components. In The Oxford Handbook of Event-Related Potential Components (pp. 329–360). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195374148.001.0001

Kappenman, E. S., & Luck, S. J. (2012b). ERP components: The ups and downs of brainwave recordings. In The Oxford handbook of event-related potential components. (pp. 3–30). Oxford University Press.

Kiss, M., Van Velzen, J., & Eimer, M. (2008). The N2pc component and its links to attention shifts and spatially selective visual processing. Psychophysiology, 45(2), 240–249. https://doi.org/10.1111/j.1469-8986.2007.00611.x

Kroenke, C. D., & Bayly, P. V. (2018). How forces fold the cerebral cortex. Journal of Neuroscience, 38(4), 767–775. https://doi.org/10.1523/JNEUROSCI.1105-17.2017

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology, 62, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4(NOV). https://doi.org/10.3389/fpsyg.2013.00863

Li, D., Hu, Y., Qi, M., Zhao, C., Jensen, O., Huang, J., & Song, Y. (2023). Prioritizing flexible working memory representations through retrospective attentional strengthening. NeuroImage, 269. https://doi.org/10.1016/j.neuroimage.2023.119902

Li, Y., Zhang, M., Liu, S., & Luo, W. (2022). EEG decoding of multidimensional information from emotional faces. NeuroImage, 258. https://doi.org/10.1016/j.neuroimage.2022.119374

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. Frontiers in Human Neuroscience, 8(1 APR), 1–14. https://doi.org/10.3389/fnhum.2014.00213

Luck, S. J. (2014). An Introduction To The Event-Related Potential Technique (2nd Editio). MIT Press Journals.

Luck, S. J. (2023). ERPLAB Decoding Tutorial. https://github.com/ucdavis/erplab/wiki/ERPLAB-Decoding-Tutorial

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods, 164(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

MathWorks. (2023). MATLAB version: 23.2.0 (R2023b). The MathWorks Inc. https://www.mathworks.com

Mazza, V., Turatto, M., Umiltà, C., & Eimer, M. (2007). Attentional selection and identification of visual objects are reflected by distinct electrophysiological responses. Experimental Brain Research, 181(3), 531–536. https://doi.org/10.1007/s00221-007-1002-4

Murphy, B., Poesio, M., Bovolo, F., Bruzzone, L., Dalponte, M., & Lakany, H. (2011). EEG decoding of semantic category reveals distributed representations for single concepts. Brain and Language, 117(1), 12–22. https://doi.org/10.1016/j.bandl.2010.09.013

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. In Clinical Neurophysiology (Vol. 118, Issue 12, pp. 2544–2590). https://doi.org/10.1016/j.clinph.2007.04.026

Nadra, J. G., Bengson, J. J., Morales, A. B., & Mangun, G. R. (2023). Attention without Constraint: Alpha Lateralization in Uncued Willed Attention. ENeuro, 10(6). https://doi.org/10.1523/ENEURO.0258-22.2023

Nemrodov, D., Niemeier, M., Patel, A., & Nestor, A. (2018). The neural dynamics of facial identity processing: Insights from EEG-based pattern analysis and image reconstruction. ENeuro, 5(1). https://doi.org/10.1523/ENEURO.0358-17.2018

Noah, S., Powell, T., Khodayari, N., Olivan, D., Ding, M., & Mangun, G. R. (2020). Neural Mechanisms of Attentional Control for Objects: Decoding EEG Alpha When Anticipating Faces, Scenes, and Tools. Journal of Neuroscience, 40(25), 4913–4924. https://doi.org/10.1523/JNEUROSCI.2685-19.2020

Polich, J. (2012). Neuropsychology of P300. The Oxford Handbook of Event-Related Potential Components, 159–188.

Prime, D. J., & Jolicoeur, P. (2010). Mental Rotation Requires Visual Short-term Memory: Evidence from Human Electric Cortical Activity. Journal of Cognitive Neuroscience. https://doi.org/https://doi.org/10.1162/jocn.2009.21337

R Core Team. (2023). R: A Language and Environment for Statistical Computing. https://www.R-project.org/

Rossion, B., & Jacques, C. (2011). The N170: Understanding the time course of face perception in the human brain. The Oxford Handbook of ERP Components, 115–142.

Sawaki, R., Geng, J. J., & Luck, S. J. (2012). A common neural mechanism for preventing and terminating the allocation of attention. Journal of Neuroscience, 32(31), 10725–10736. https://doi.org/10.1523/JNEUROSCI.1864-12.2012

Smith, F. W., & Smith, M. L. (2019). Decoding the dynamic representation of facial expressions of emotion in explicit and incidental tasks. NeuroImage, 195, 261–271. https://doi.org/10.1016/j.neuroimage.2019.03.065

Trammel, T., Khodayari, N., Luck, S. J., Traxler, M. J., & Swaab, T. Y. (2023). Decoding semantic relatedness and prediction from EEG: A classification method comparison. NeuroImage, 277, 120268. https://doi.org/10.1016/j.neuroimage.2023.120268

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. NeuroImage, 137, 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012

Wang, D., Miao, D., & Blohm, G. (2012). Multi-class motor imagery EEG decoding for brain-computer interfaces. Frontiers in Neuroscience, OCT. https://doi.org/10.3389/fnins.2012.00151

Wolff, M. J., Ding, J., Myers, N. E., & Stokes, M. G. (2015). Revealing hidden states in visual working memory using electroencephalography. Frontiers in Systems Neuroscience, 9(september). https://doi.org/10.3389/fnsys.2015.00123

Yang, L., Song, Y., Ma, K., & Xie, L. (2021). Motor Imagery EEG Decoding Method Based on a Discriminative Feature Learning Strategy. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29, 368–379. https://doi.org/10.1109/TNSRE.2021.3051958

Zhang, W., & Kappenman, E. S. (2023). Maximizing signal-to-noise ratio and statistical power in ERP measurement: Single sites versus multi-site average clusters. Psychophysiology. https://doi.org/10.1111/psyp.14440

Supplementary Materials for

Using Multivariate Pattern Analysis to Increase Effect Sizes for Event-Related Potential Analyses

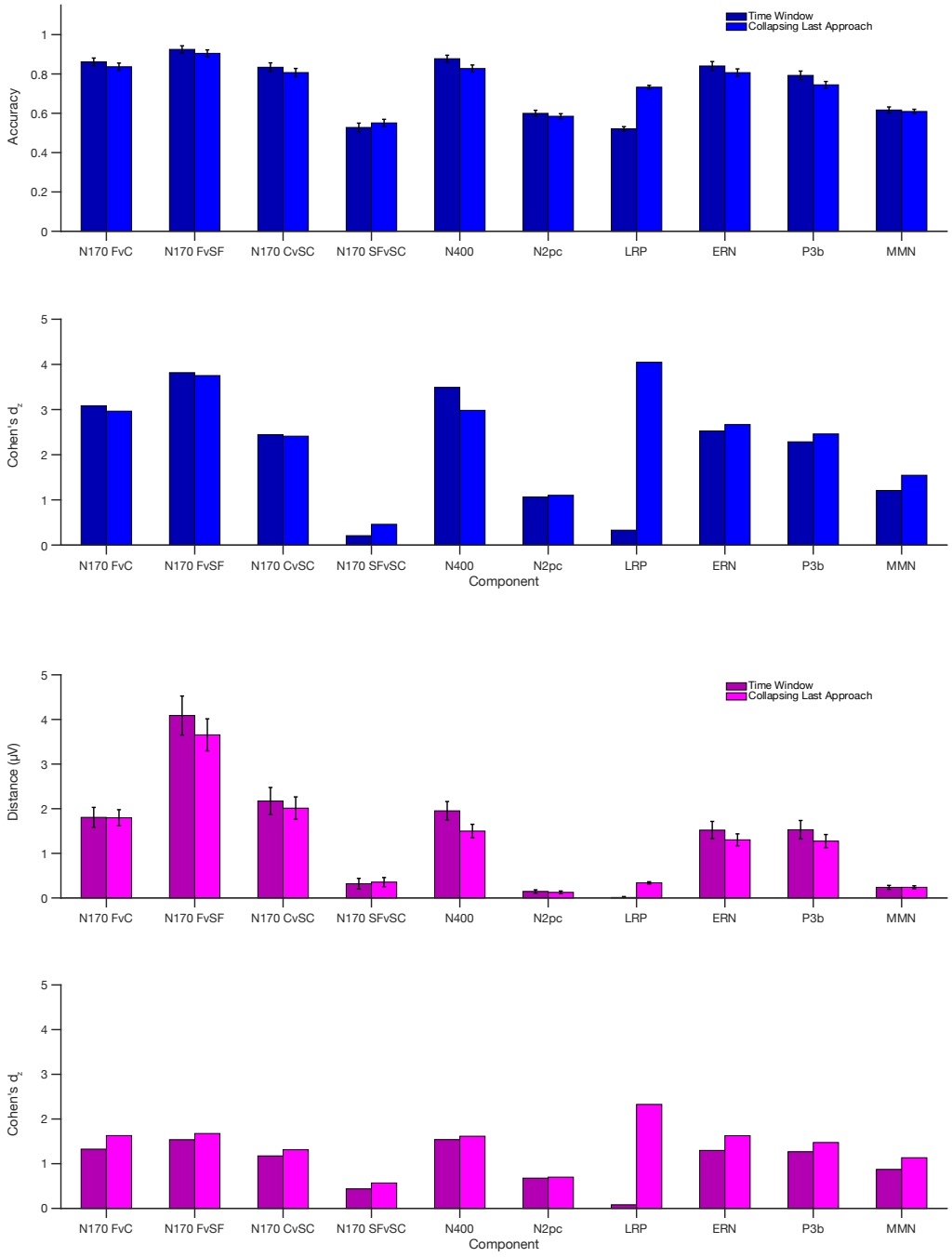


Figure S1. Comparison of two approaches for computing time-window values for the SVM and crossnobis methods. In the collapsing-last approach, we simply performed decoding on the individual time points and averaged the point-by-point decoding accuracy values and crossnobis distances across the measurement window. In the collapsing-first approach, we averaged the voltages across the time window in each single-trial epoch prior to the decoding and crossnobis analyses.

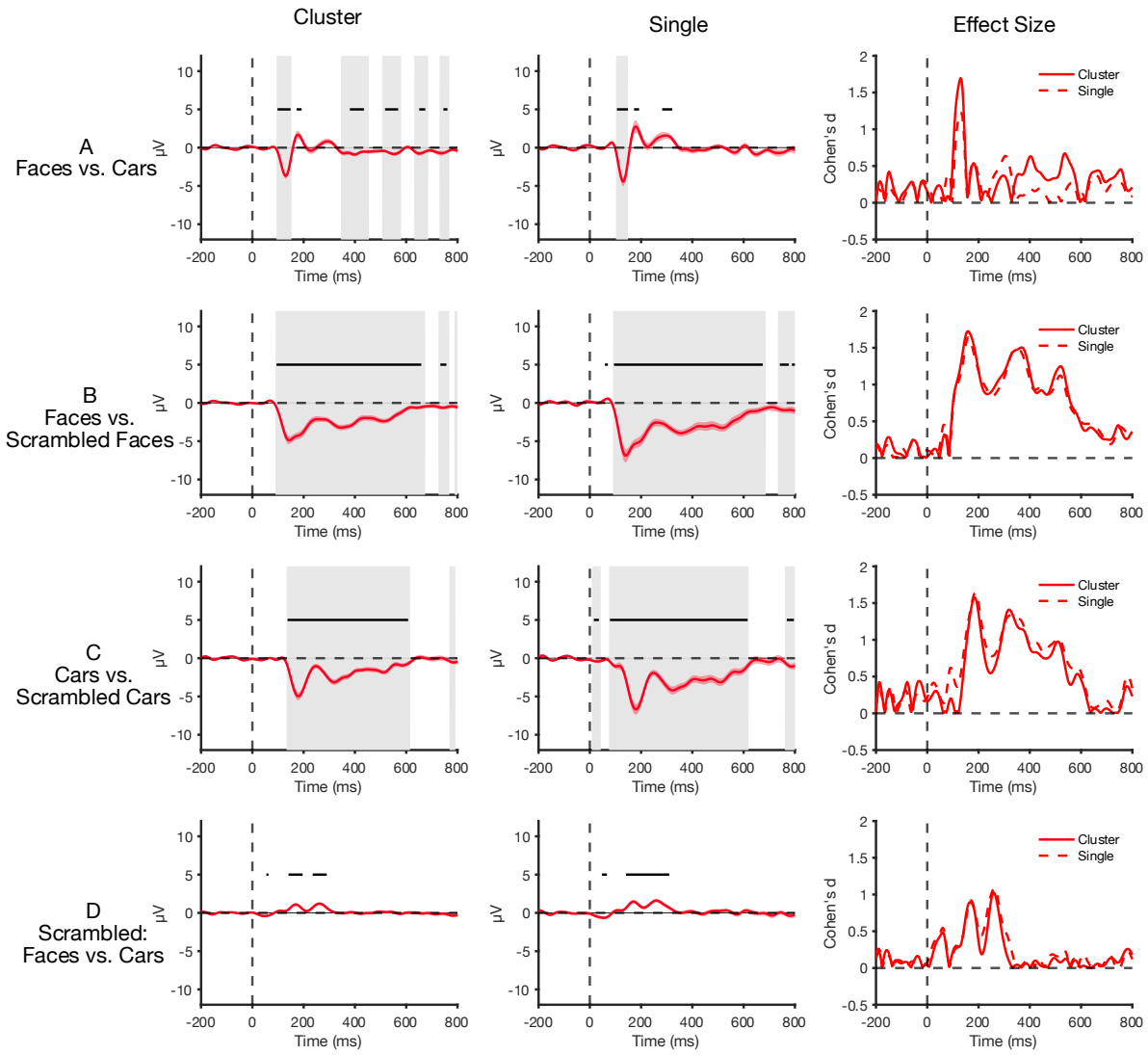


Figure S2. Point-by-point ERP difference voltages, for the 4 components elicited by the N170 paradigm using single sites, and clusters of sites. The shaded bands around the waveform show the standard error of the mean. Gray shading indicates significance at $p < 0.05$, FDR corrected. The right column shows the effect size (Cohen's d_z) at each time point for each analysis method. The cluster sites tend to yield larger effect sizes than using single sites.

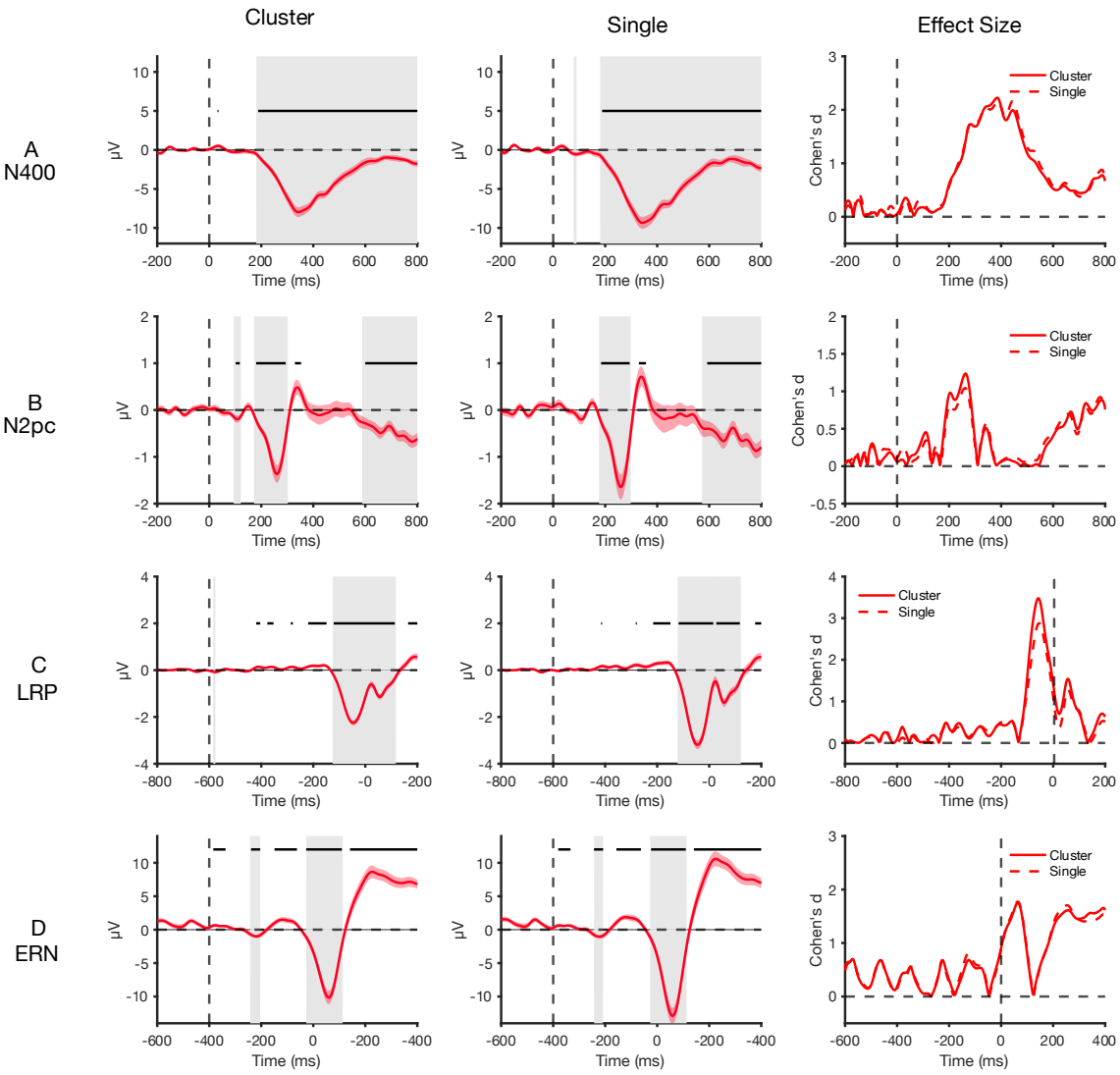


Figure S3. Point-by-point ERP difference voltages, for the 4 components elicited by the N400, N2pc, LRP/ERN paradigms using single sites and clusters of sites. The shaded bands around the waveform show the standard error of the mean. Gray shading indicates significance at $p < 0.05$, FDR corrected. The right column shows the effect size (Cohen's d_z) at each time point for each analysis method. The cluster sites tend to yield larger effect sizes than using single sites.

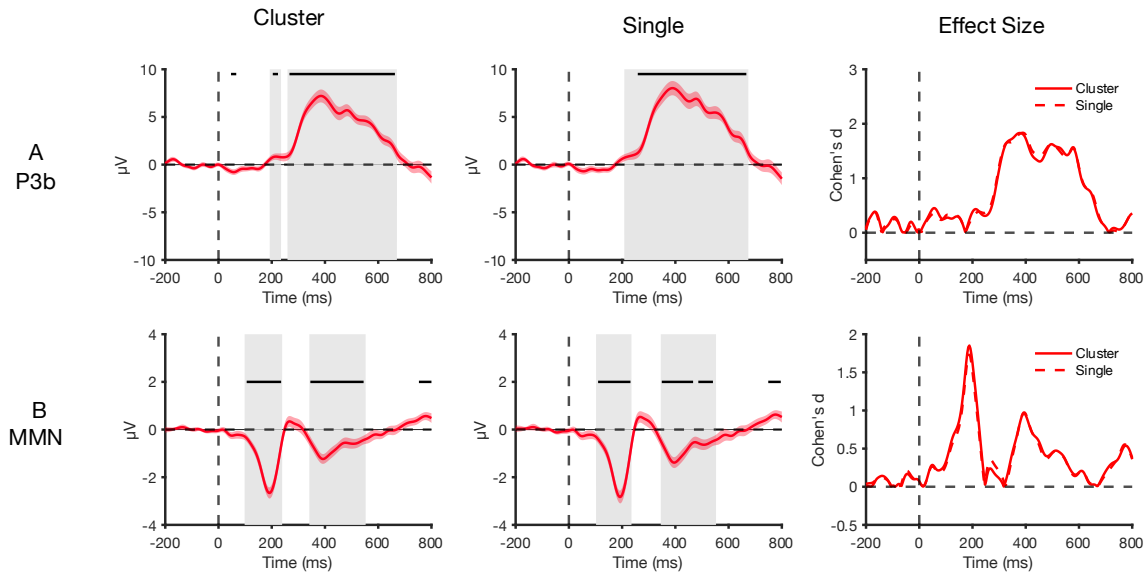


Figure S4. Point-by-point ERP difference voltages, for the components elicited by the oddball paradigms (P3b/MMN) using single sites and clusters of sites. The shaded bands around the waveform show the standard error of the mean. Gray shading indicates significance at $p < 0.05$, FDR corrected. The right column shows the effect size (Cohen's d_z) at each time point for each analysis method. The cluster sites tend to yield roughly the same effect sizes as their single site counterparts.

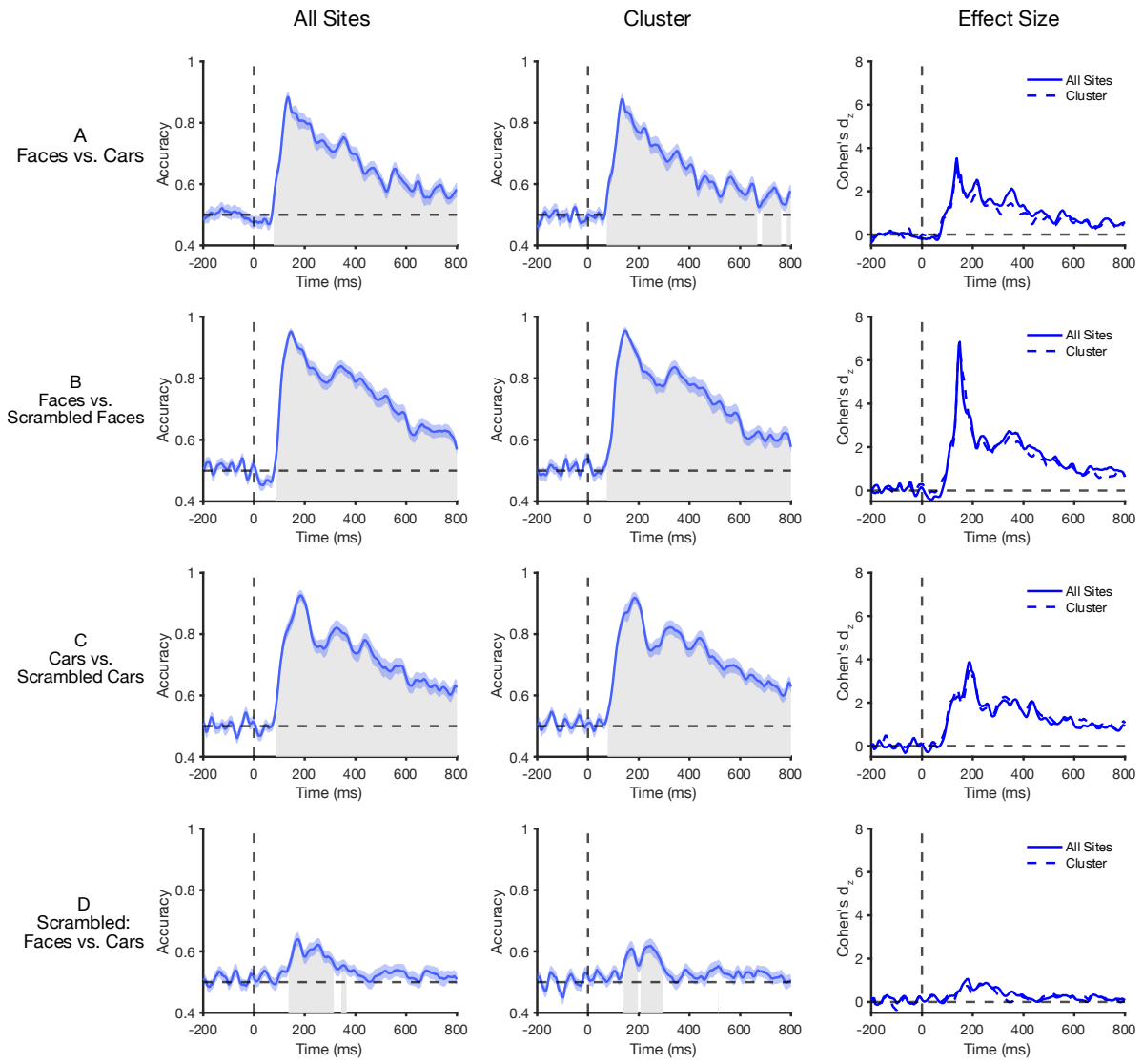


Figure S5. Point-by-point decoding accuracy for the 4 comparisons analogous to the components elicited by the N170 paradigm using clusters of sites, and all sites. The shaded bands around the waveform show the standard error of the mean. Gray shading indicates significance at $p < 0.05$, FDR corrected. The right column shows the effect size (Cohen's d_z) at each time point for each analysis method. Using all sites tends to yield larger effect sizes than using a cluster.

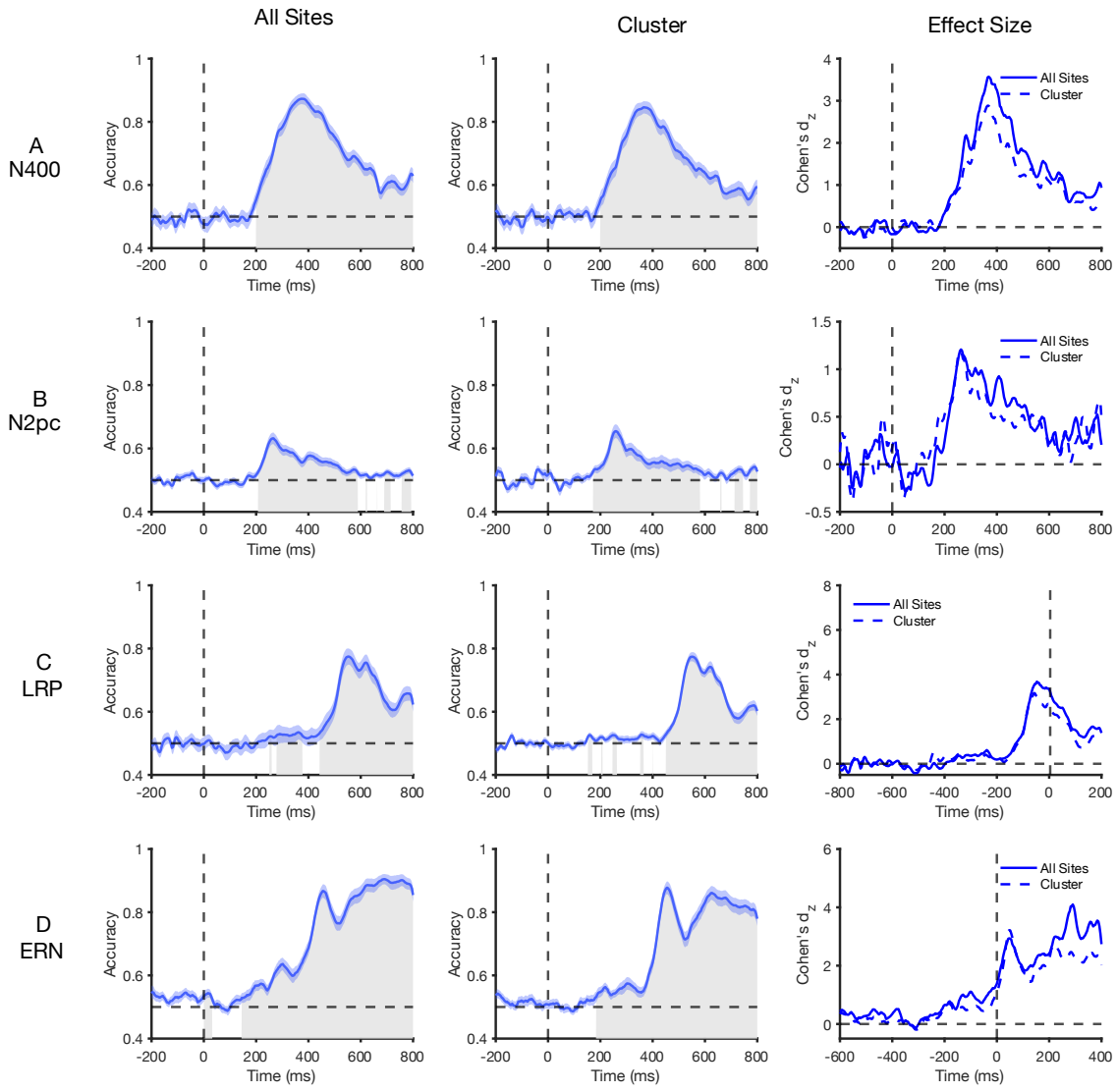


Figure S6. Point-by-point decoding accuracy for the 4 comparisons analogous to the components elicited by the N400, N2pc, LRP/ERN paradigms using clusters of sites, and all sites. The shaded bands around the waveform show the standard error of the mean. Gray shading indicates significance at $p < 0.05$, FDR corrected. The right column shows the effect size (Cohen's d_z) at each time point for each analysis method. Using all sites tends to yield larger effect sizes than using a cluster.

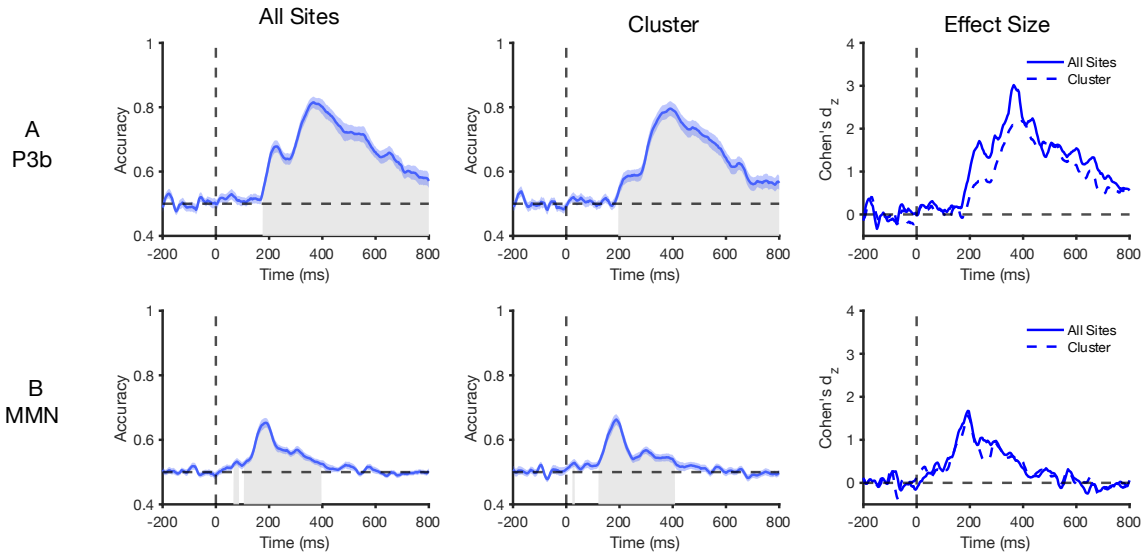


Figure S7. Point-by-point decoding accuracy for the comparisons analogous to the components elicited by the P3b and MMN paradigms using clusters of sites, and all sites. The shaded bands around the waveform show the standard error of the mean. Gray shading indicates significance at $p < 0.05$, FDR corrected. The right column shows the effect size (Cohen's d_z) at each time point for each analysis method. Using all sites tends to yield roughly the same effect sizes as using single sites.

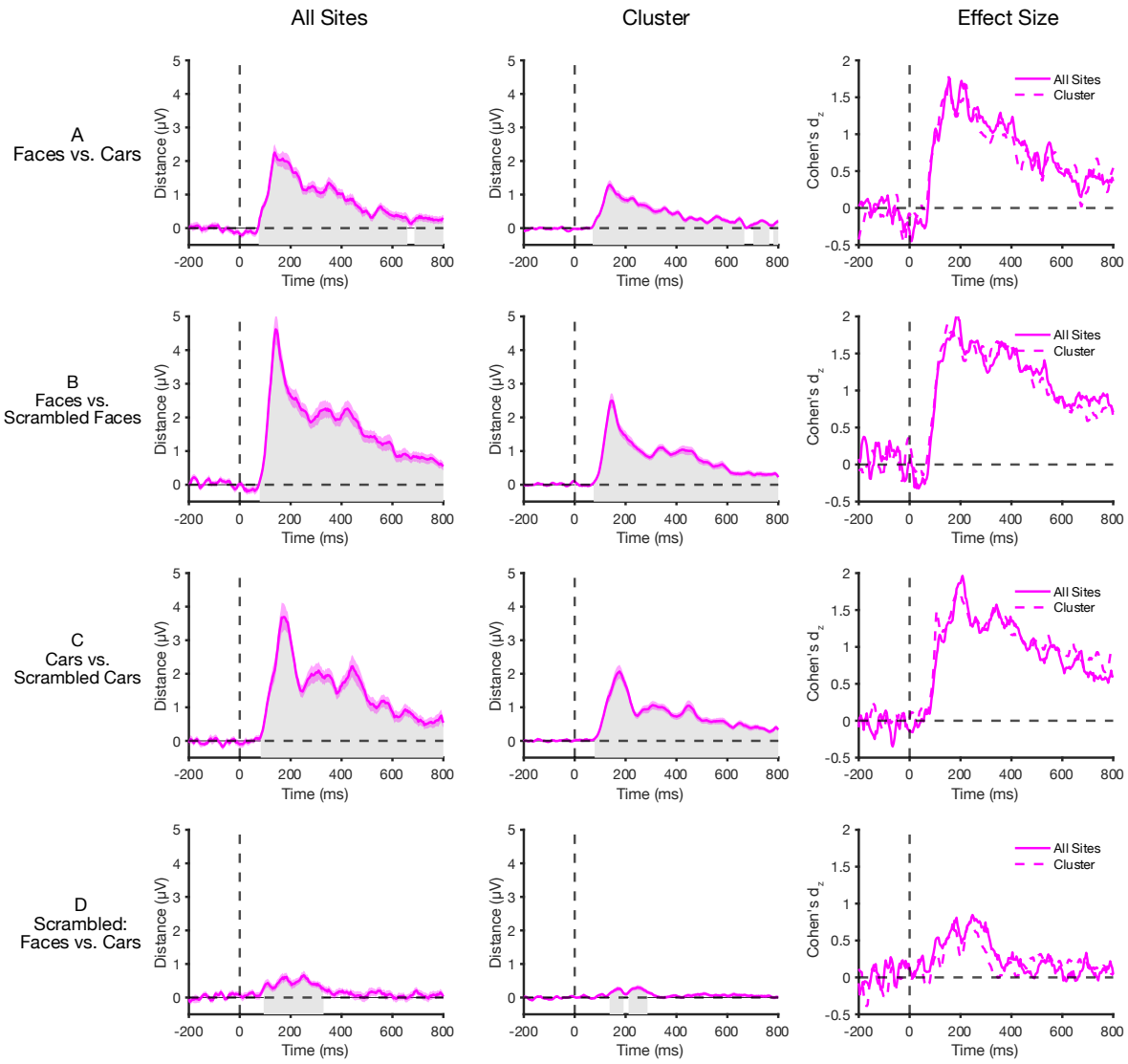


Figure S8. Point-by-point crossnobis distance for the 4 comparisons analogous to the components elicited by the N170 paradigm using clusters of sites, and all sites. The shaded bands around the waveform show the standard error of the mean. Gray shading indicates significance at $p < 0.05$, FDR corrected. The right column shows the effect size (Cohen's d_z) at each time point for each analysis method. Using all sites or clusters tend to yield roughly the same effect sizes across the time series.

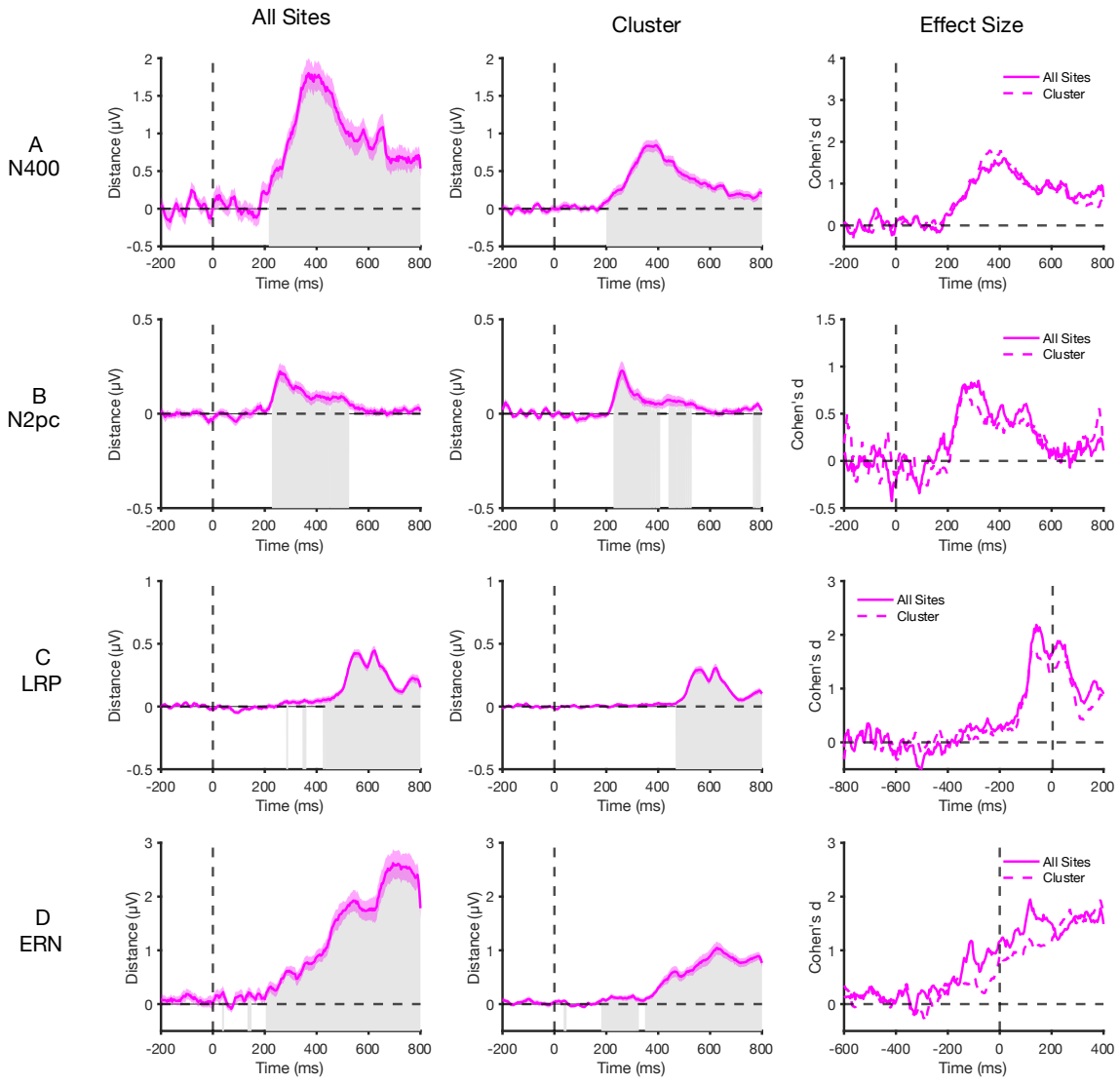


Figure S9. Point-by-point crossnobis distance for the 4 comparisons analogous to the components elicited by the N400, N2pc, LRP/ERN paradigms using clusters of sites, and all sites. The shaded bands around the waveform show the standard error of the mean. Gray shading indicates significance at $p < 0.05$, FDR corrected. The right column shows the effect size (Cohen's d_z) at each time point for each analysis method. Using all sites or clusters tend to yield roughly the same effect sizes across the time series.

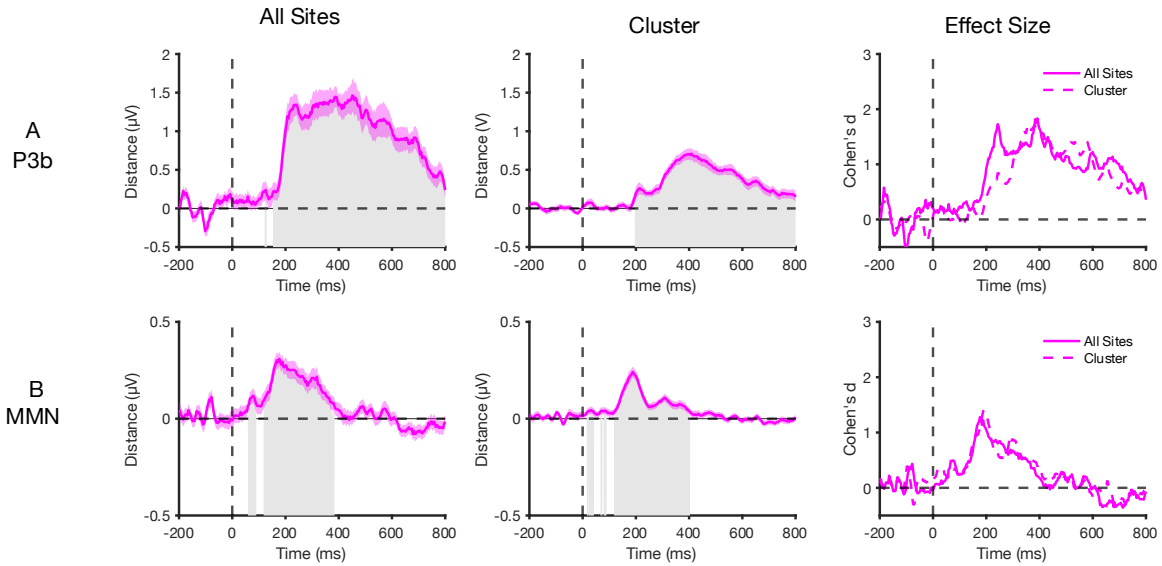


Figure S10. Point-by-point crossnobis distance for the comparisons analogous to the components elicited by the P3b and MMN paradigms using clusters of sites, and all sites. The shaded bands around the waveform show the standard error of the mean. Gray shading indicates significance at $p < 0.05$, FDR corrected. The right column shows the effect size (Cohen's d_z) at each time point for each analysis method. Using all sites tends to yield roughly the same effect sizes as using single sites.

Table S1. Summary of univariate analyses when the data were measured from a single channel instead of a cluster of channels.

Univariate Parameters

ERP Component	Measurement Window	Electrode Measured	N	df	Mean Amp t	Mean Amp p	Time Locking Point
N170 FvC	110 to 150	PO8	37	36	7.36	$p < 0.0001$	stimulus onset
N170 FvSF	110 to 150	PO8	37	36	7.90	$p < 0.0001$	stimulus onset
N170 CvSC	110 to 150	PO8	37	36	4.38	$p < 0.0001$	stimulus onset
N170 SFvSC	110 to 150	PO8	37	36	3.13	$p = 0.004$	stimulus onset
N400	300 to 500	CPz	37	36	13.96	$p < 0.0001$	stimulus onset
N2pc	200 to 275	PO7/PO8	35	34	5.86	$p < 0.0001$	stimulus onset
LRP	-100 to 0	C3/C4	37	36	15.56	$p < 0.0001$	response onset
ERN	0 to 100	FCz	36	35	9.44	$p < 0.0001$	response onset
P3	300 to 600	Pz	34	33	10.82	$p < 0.0001$	stimulus onset
MMN	125 to 225	FCz	39	38	9.49	$p < 0.0001$	stimulus onset

Table S2. Summary of SVM results when the decoding was performed only on the data from the electrodes used in the main univariate analyses

			SVM Parameters								
ERP Component	Measurement Window	Electrode Sites	Avg # of Decoding Trials	Std Deviation	# of Crossfolds	Min # of Trials Per Crossfold	N	df	Decoding t	Decoding p	Time Locking Point
N170 FvC	110 to 150	P7, P8, PO7, PO8, O1,O2	68.41	6.95	5	13	37	36	41.00	p < 0.0001	stimulus onset
N170 FvSF	110 to 150	P7, P8, PO7, PO8, O1,O2	67.43	6.87	5	13	37	36	46.81	p < 0.0001	stimulus onset
N170 CvSC	110 to 150	P7, P8, PO7, PO8, O1,O2	67.16	8.05	5	13	37	36	40.36	p < 0.0001	stimulus onset
N170 SFvSC	110 to 150	P7, P8, PO7, PO8, O1,O2	66.65	7.63	5	13	37	36	32.58	p < 0.0001	stimulus onset
N400	300 to 500	Cz, Pz, C3, C4, P3, P4	51.05	5.08	4	10	37	36	38.90	p < 0.0001	stimulus onset
N2pc	200 to 275	P7/8, PO7/8, PO3/4, O1/2, P3/4	57.54	9.87	5	11	35	34	6.68	p = 0.0002	stimulus onset
LRP	-100 to 0	F3/4, FC3/4, C3/4, C5/6, P3/4	84.15	7.41	5	16	37	36	19.42	p < 0.0001	response onset
ERN	0 to 100	Fz, FCz, Cz,	41.17	21.12	4	10	36	35	41.82	p < 0.0001	response onset

ERP Component	Measurement Window	Electrode Sites	Avg # of Decoding Trials	Std Deviation	# of Crossfolds	Min # of Trials Per Crossfold	N	df	Decoding t	Decoding p	Time Locking Point
P3	300 to 600	FC3, FC4 Cz, CPz, Pz, P3, P4	33.62	4.05	3	11	34	33	44.51	p < 0.0001	stimulus onset
MMN	125 to 225	Fz, FCz, Cz, FC3, FC4	185.77	13.94	10	18	39	38	49.95	p < 0.0001	stimulus onset

Supplementary Table 3. Summary of crossnobis results when the decoding was performed only on the data from the electrodes used in the main univariate analyses.

Crossnobis Parameters

ERP Component	Measurement Window	Electrode Sites	N	df	Xnobis t	Xnobis p	Time Locking Point
N170 FvC	110 to 150	P7, P8, PO7, PO8, O1,O2	37	36	10.59	p < 0.0001	stimulus onset
N170 FvSF	110 to 150	P7, P8, PO7, PO8, O1,O2	37	36	11.60	p < 0.0001	stimulus onset
N170 CvSC	110 to 150	P7, P8, PO7, PO8, O1,O2	37	36	8.37	p < 0.0001	stimulus onset
N170 SFvSC	110 to 150	P7, P8, PO7, PO8, O1,O2	37	36	2.47	p = 0.008	stimulus onset
N400	300 to 500	Cz, Pz, C3, C4, P3, P4	37	36	11.19	p < 0.0001	stimulus onset
N2pc	200 to 275	P7/8, PO7/8, PO3/4, O1/2, P3/4	35	34	4.14	p < 0.0001	stimulus onset

ERP Component	Measurement Window	Electrode Sites	N	df	Xnobis t	Xnobis p	Time Locking Point
LRP	-100 to 0	F3/4, FC3/4, C3/4, C5/6, P3/4	37	36	11.14	$p < 0.0001$	response onset
ERN	0 to 100	Fz, FCz, Cz, FC3, FC4	36	35	6.58	$p < 0.0001$	response onset
P3	300 to 600	Cz, CPz, Pz, P3, P4	34	33	11.33	$p < 0.0001$	stimulus onset
MMN	125 to 225	Fz, FCz, Cz, FC3, FC4	39	38	8.58	$p < 0.0001$	stimulus onset

Chapter 3: Enhanced Working Memory Representations for Rare Events

Carlos Daniel Carrasco, Aaron Matthew Simmons, John E. Kiat & Steven J. Luck

Center for Mind & Brain, University of California, Davis

*This work is currently being prepped for submission:

A preprint is available here: https://doi.org/10.1101/2024.03.20.585952

CC-By Attribution 4.0 International, https://creativecommons.org/licenses/by/4.0/.

Abstract

Rare events (oddballs) produce a variety of enhanced physiological responses relative to frequent events (standards), including the P3b component of the event-related potential (ERP) waveform. Previous research has suggested that the P3b component is related to working memory, which implies that working memory representations will be enhanced for rare stimuli. To test this hypothesis, we devised a modified oddball paradigm in which the target was a disk presented at one of 16 different locations, which were divided into a rare set and a frequent set. Participants made a binary response on each trial to report whether the target appeared in the rare set or the frequent set. As expected, the P3b was much larger for stimuli appearing at a location within the rare set. We also included occasional probe trials in which the subject reported the exact location of the target. We found that these reports were more accurate for locations within the rare set than for locations within the frequent set. Moreover, the mean accuracy of these reports was correlated with the mean amplitude of the P3b. We also applied multivariate pattern analysis to the ERP data to "decode" the remembered location of the target. Decoding accuracy was greater for locations within the rare set than for locations within the frequent set. These behavioral and electrophysiological results demonstrate that although both frequent and rare events are stored in working memory, the representations are enhanced for rare events.

Significance Statement

For many decades, researchers have observed that rare events elicit a broad range of physiological responses, and there has been much speculation about the functional significance of these responses. One such response is the P3b component, which is a large voltage deflection in scalp EEG recordings. Over 40 years ago, the P3b was hypothesized to reflect "context updating" (now often called "working memory updating"). However, there has been no direct evidence that working memory is actually enhanced for rare, P3b-eliciting events. In the present study, we found that both behavioral and electrophysiological measures of working memory were enhanced for rare events. This is potentially related to the release of norepinephrine across the cortex.

Introduction

Rare task-relevant events (often called oddballs) generate a variety of physiological responses, including increased firing of noradrenergic neurons in the locus coeruleus, phasic pupil dilation, increased blood-oxygen-level dependent (BOLD) activity in a variety of cortical areas, and a large P3b event-related potential (ERP) component (Bledowski et al., 2004; Clark et al., 2000; Johnson, 1988; Krebs et al., 2018; Linden et al., 1999; Murphy et al., 2011; Polich, 1986; Soltani & Knight, 2000). There has been considerable speculation about the functional significance of these physiological changes (Kim, 2014; Linden et al., 1999; Nieuwenhuis et al., 2011; Paller et al., 1992). A common hypothesis is that the P3b activity elicited by oddballs is related to working memory encoding, although the details vary across theories (Donchin & Coles, 1988; Kok, 2001; Polich, 2007, 2012). Moreover, oddballs elicit phasic increases in attention (Aston-Jones & Cohen, 2005; Katayama & Polich, 1998; Kim, 2014; Murphy et al., 2011), which might also enhance working memory. However, we know of no direct evidence that working memory is actually enhanced for relatively rare events compared to relatively frequent events. The goal of the present study was to test this hypothesis, using both behavioral and electrophysiological measures.

Typical oddball paradigms do not provide a sensitive assessment of working memory. For example, a typical paradigm would involve presenting a sequence of stimuli in which 90% are the letter X and 10% are the letter O, and the task would be to press one button for Xs and another button for Os. The responses are made immediately, so it is not necessary to store the Xs and Os in working memory. Moreover, the Xs and Os are so easily discriminable that memory performance would likely be at ceiling if tested after a brief delay.

We therefore developed the modified oddball paradigm shown in Figure 1. On each trial, a target disc appeared briefly near one of eight locations around a circle, and the main task was to press one button if the disc appeared near one of the cardinal axes (up, down, left, or right) and a different button if it appeared near one of the diagonal axes (upper left, upper right, lower left, or lower right). One of these two categories was rare (12.5%), and the other was frequent (87.5%). This was much like a traditional oddball task, in which participants make an immediate response to indicate whether the stimulus belonged to the rare category or the frequent category. We assumed that stimuli belonging to the rare category (the oddballs) would elicit a larger P3b component than stimuli belonging to the frequent category (the standards). To provide a sensitive measure of the working memory representation of the disk, we also included occasional probe trials, on which participants were asked to click on the exact location of the disc from that trial after a brief delay. If working memory is enhanced for rare stimuli, then the target localization response on probe trials should be more accurate following an oddball than following a standard. Moreover, if the P3b component elicited by oddballs is associated with working memory updating, then participants with greater P3b amplitudes for the oddballs should exhibit more accurate memory on the probe trials than individuals with smaller P3b amplitudes.

In addition, we applied multivariate pattern analysis (MVPA) to the ERP data to decode the remembered location of the disc in the delay period following each stimulus. This provided a means of monitoring the working memory representation of the stimuli during the delay period following each target, independently of the decision and response processes that are involved in the behavioral responses. Specifically, we decoded which of the four locations within a category was presented (e.g., up, down, left, or right for the cardinal category). We predicted that the

within-category decoding accuracy would be greater for a given category when that category was rare than when it was frequent.

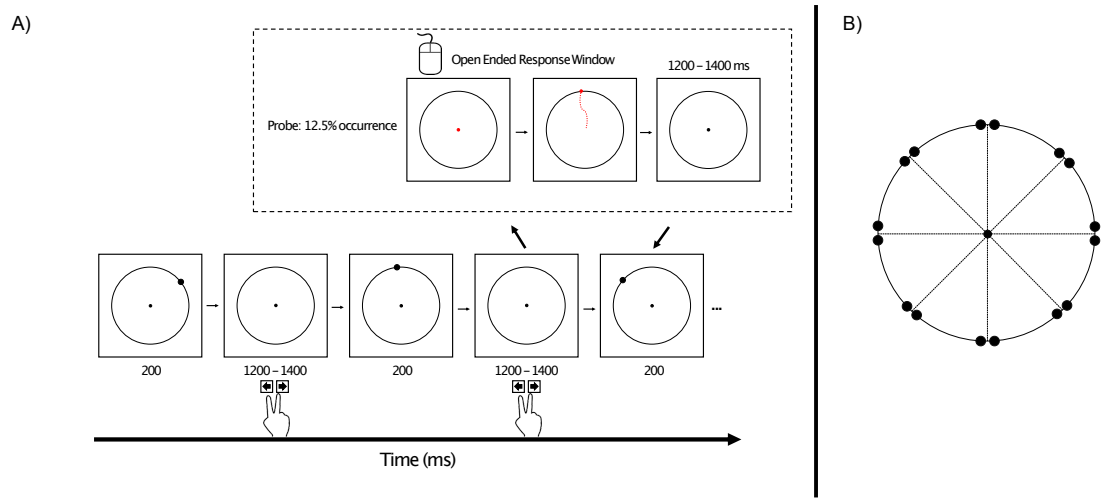


Figure 1. (A) Overview of the oddball task. On every trial, participants made a speeded keypress following the presentation of a target disc to indicate whether it appeared near a cardinal axis or near a diagonal axis (as shown in B). For a given trial block, one axis was rare and the other was frequent. On 12.5% of trials (probe trials), the fixation point turned red at the offset of the intertrial interval, which signaled participants that they should use the mouse to click on the remembered location of the target disk from that trial. (B) Possible locations of the target disc, which could appear 1° counterclockwise or 1° clockwise from a cardinal or diagonal axis. Note that the distance of the targets from the axes is slightly exaggerated in this image.

Methods

Participants

Twenty-two human participants from the UC Davis community completed the study (14 women, 7 men, 1 unreported gender; mean age = 20, SD = 1.69). All participants were neurotypical and had normal or corrected-to-normal vision with no history of neurological conditions. Twenty participants were right-handed, and two were left-handed. We chose to collect a somewhat higher N than is typical of multivariate pattern analysis EEG studies (e.g., N = 16 in Bae & Luck, 2019; Bae, 2021; Bae & Luck, 2018) because of the relatively small number of oddball trials in the present study. Monetary compensation was provided at a rate of

$15/hr. All participants provided informed consent, and the study was approved by the University of California, Davis Institutional Review Board.

Stimuli and Task

Figure 1 illustrates the stimuli and task. The experiment presentation script can be downloaded at doi.org/10.17605/OSF.IO/HV7JU.

Stimuli were presented using Psychopy (Peirce et al., 2019) on an LCD monitor (Dell U2412M) with a gray background (27.7 cd/m^2) at a viewing distance of 100 cm. A 0.05° black fixation dot was continuously visible in the center of the monitor, surrounded by a black circle with a radius of 2.17°.

On each regular trial, a black target disc (0.2°) appeared for 200 ms, followed by an interstimulus interval of 1200-1400 ms (rectangular distribution). The target appeared on the black circle, near one of the cardinal axes (0°, 90°, 180°, 270°) or near one of the diagonal axes (45°, 135°, 225°, 315°), centered either -1 or +1 degrees from one of these axes (e.g., 44°, 46°, 89°, 91°). Thus, there were 16 possible target locations, 8 near the cardinal axes and 8 near the diagonal axes. Placing the target ±1° from an axis was designed to require participants to remember the precise location of the target rather than relying on simple categories such as “top” or “lower left”. In addition, this made it possible to examine biases away from the cardinal axes (Bae, 2022).

For half of the trial blocks, the target appeared near a cardinal axis on 12.5% of trials (oddballs) and appeared near a diagonal axis on the remaining 87.5% of trials (standards). This was reversed for the other trial blocks. Each of the 8 locations within the cardinal or diagonal category occurred with equal probability. Participants were instructed to make a speeded response on a computer keyboard to indicate whether the immediately preceding target was near

a cardinal axis or near a diagonal axis. Half the participants pressed the left arrow key for cardinal and the right arrow key for diagonal; this was reversed for the other half.

Although most oddball studies use only one rare stimulus (e.g., a high-pitched tone) and one frequent stimulus (e.g., a low-pitched tone), the P3b component is sensitive to the probability of the task-defined category rather than the probability of the physical stimulus (Luck, 2014; Mecklinger & Ullsperger, 1993). Oddball studies have used abstract categories such as the ones used here for decades (e.g., Kutas et al., 1977).

Each block contained 256 trials (32 oddballs and 224 standards). The blocks alternated between cardinal-oddball/diagonal-standard and diagonal-oddball/cardinal-standard, with the starting condition counterbalanced across participants. Each participant received 256 oddballs and 1792 standards, with each of the 16 locations occurring equally often within each of these categories (2048 trials total). Thus, there were 112 trials for a given location when that location was in the standard category and 16 trials for a given location when that location was in the oddball category.

Working memory for the exact target location was probed following a random 12.5% of targets. When a target was probed, the fixation dot changed from black to red at the end of the interstimulus interval (i.e., after the participant indicated whether the target had been near a cardinal or diagonal axis). Once it turned red, the participant could use the mouse to move the red dot. They were instructed to move the red dot to the remembered location of the target and then click the mouse button (with no time pressure). The red dot then disappeared and was replaced by the black fixation dot in the center of the monitor. The stream of targets then resumed after a delay of 1200-1400 ms. For each participant, oddballs were probed on 32 trials and standards were probed on 224 trials.

Note that participants could not predict whether a given trial would be a probe trial. Probe trials were just like any other trials until the end of the intertrial interval following a given target. Thus, the working memory representations and neural activity during the intertrial interval following a target could not be systematically different on probe and non-probe trials.

We probed on only a small subset of trials for two reasons. First, we wanted the task to be more like a traditional oddball task, in which exact memories are not probed. Second, probe responses took considerable time, and we could obtain more targets in a session of a reasonable duration if we probed infrequently. A very large overall number of trials was needed to obtain a sufficient number of trials per location for the EEG decoding (which used the data from all trials, not just probe trials). Far fewer trials were needed to obtain robust measures of behavioral accuracy.

EEG Recording and Preprocessing

The continuous EEG was recorded using a Brain Products actiCHamp recording system. We recorded EEG signals from 27 standard 10/20 sites: FP1, Fz, F3, F7, Cz, C3, Pz, P3, P5, P7, P9, PO7, PO3, O1, POz, Oz, FP2, F4, F8, C4, P4, P6, P8, P10, PO4, PO8, O2. We also recorded signals from horizontal electrooculogram (HEOG) electrodes lateral to the left and right external canthi, from a vertical electrooculogram (VEOG) electrode located under the right eye, and from electrodes over the left and right mastoids. Single-ended voltages were recorded relative to a ground electrode located at AFz. All electrode impedances were kept < 50 KΩ. The signals were filtered with a cascaded integrator-comb antialiasing filter (half-power cutoff at 130 Hz) and digitized at 500 Hz.

All preprocessing steps were conducted in MATLAB using the EEGLAB and ERPLAB toolboxes (Delorme & Makeig, 2004; Lopez-Calderon & Luck, 2014), following the standard

pipeline described by Luck (2023).We began by shifting the event codes to account for the monitor delay (56 ms, which was measured using a photodiode). The signals were then resampled at 250 Hz. The DC offset was removed, and the signals were high-pass filtered (noncausal Butterworth impulse response function, half-amplitude cutoffs at 0.1 Hz, 12 dB/oct roll-off). Time segments between the trial blocks were deleted, and the EEG data were referenced to the P9 electrode site. A bipolar VEOG channel was created by subtracting the FP2 electrode from the VEOG electrode, and a bipolar HEOG channel was created by subtracting HEOG-right from HEOG-left.

Independent component analysis (ICA) was then performed to correct for blinks and eye movements (excluding the bipolar channels). The data used for the ICA decomposition were filtered more aggressively (noncausal Butterworth impulse response function, half-amplitude cutoffs at 1 – 30 Hz, 48 dB/oct roll-off) and resampled at 100 Hz. The ICA weights were then transferred back to the original data, and independent components corresponding to blinks and artifacts were removed from the data; typically, 1-2 components were removed per participant. We used consistency between the shape, timing and spatial location of the component compared to the bipolar HEOG and VEOG signals to determine which components were artifacts.

The ICA-corrected data were then re-referenced to the average of the left and right mastoid electrodes. The data were then segmented from -500 to 1496 ms relative to stimulus onset and baseline-corrected to the mean voltage from -500 to 0 ms. Finally, epochs were marked for rejection using standard ERPLAB routines if they contained large voltage deflections in any channel (using a simple voltage threshold) or an eyeblink that would have prevented perception of the target (using a moving window peak-to-peak function from 0 to 200 ms in the bipolar VEOG channel). This led to an average rejection of 13.7% of trials across participants

(SE = 6.4%). We always exclude participants for whom more than 25% of trials were rejected; no participants exceeded this threshold in the present study.

Behavioral Analyses

For the oddball categorization task, we computed the proportion correct and the mean response time (RT) for each participant, separately for the oddball and standard categories.

For the probe task, the stimuli and responses were coded in terms of their angular position around the circle of possible target locations. We excluded trials on which the reported location was > 40° away from the true location (0.89% ± 0.31% of trials), because such large errors presumably reflect lapses of attention. On the remaining trials, we computed the response error, defined as the angular distance between the true target location and the reported location. In the primary analyses, we took the absolute value of the response error on each trial and averaged across trials for a given participant, separately for the oddball and standard categories.

Working memory representations tend to be biased away from the cardinal axes (Bae, 2022) and we examined these biases in a separate analysis of the probe data. Following prior research, we expected that locations that were 1° clockwise from a cardinal axis would be reported as being more than 1° clockwise, and that locations that were 1° counterclockwise from a cardinal axis would be reported as being more than 1° counterclockwise. If working memory representations are enhanced for oddballs compared to standards, then these biases should be reduced for the oddballs. To analyze the biases, we took the response error on each trial and gave it a negative sign if it was farther away from the nearest cardinal axis than the true stimulus location (e.g., given a true location of 89°, which was near the cardinal axis at 90°, a reported location at 87° was coded as a response error of -2°). Similarly, we gave a response error a positive sign if it was biased in the opposite direction (e.g., given a true location of 89°, a

reported location at 91° was coded as a response error of +2°). These values were then averaged across trials for a given participant, separately for oddball and standard trials. A consistent repulsion away from the nearest axis would lead to a negative average value. Little or no bias of this sort would be expected for locations near the diagonal axes for either oddballs or standards, so trials with stimuli near the diagonal axes were excluded from this analysis.

P3b Scoring

We measured P3b amplitude at an a priori electrode size (Pz) and an a priori time window which we defined as defined as ±150 ms around the P3b peak of the grand averaged ERP. This peak was at 512 ms so our measurement window was from 362 to 662 ms. The P3b amplitude was scored as the mean voltage during this time window at Pz, separately for oddballs and standards.

Decoding Analysis

The decoding analysis collapsed across the two locations that were ±1° from a given axis (e.g., 44° and 46°), which were too similar to be reliably differentiated by the decoding process. This gave us four cardinal locations (0°, 90°, 180°, 270°) and four diagonal locations (45°, 135°, 225°, 315°). We performed the decoding separately for these two categories, separately, when a category was the oddball and when it was the standard. For example, we decoded which of the four cardinal locations was present on the cardinal trials when the cardinal category was the oddball, and we separately decoded which of the four cardinal locations was present on the cardinal trials when the cardinal category was the standard. This allowed us to compare the decoding accuracy for the same stimulus locations when those locations were oddballs and when they were standards. Decoding was performed separately at each time point (every 4 ms from -500 to +1496 ms).

We followed the location decoding procedure developed by Bae and Luck (2018), as implemented in ERPLAB Toolbox (Lopez-Calderon & Luck, 2014) using MATLAB's fitcecoc() function. The first step was to apply a 6 Hz lowpass filter (48 dB/octave roll-off), which minimized contamination from alpha-band EEG oscillations. This filter reduced the temporal resolution of the analysis, but that was not a major problem given that we were examining long-duration working memory effects. The ocular channels were left out of the analysis, leaving 27 scalp channels.

Decoding was performed on averaged ERP waveforms using support vector machines (SVM) with error-correcting output codes (Dietterich & Bakiri, 1995) and 3-fold cross-validation. We conducted four separate decoding runs for each participant: a) cardinal oddball; b) cardinal standard; c) diagonal oddball; and d) diagonal standard. For each run, there were four possible stimulus locations (the four different cardinal locations or the four different diagonal locations). The decoder attempted to determine which of these four locations was presented on the basis of the pattern of voltage across electrode sites at a given time point.

Because decoding accuracy tends to increase when more trials are available, we randomly subsampled a subset of trials from the standards to equate the number of trials for oddballs and standards. This yielded an average across participants of 19.5 trials (SD = 4.5) at each of the four locations for each decoding run. The available trials were randomly divided into three subsets, and a separate averaged ERP waveform was created from each of these three subsets. This gave us a matrix of 4 classes (4 locations) × 3 averages × 27 electrode sites × 500 time points for each participant for each decoding run. Decoding was performed on the matrix of classes × averages × electrode sites, separately at each time point for each participant for a given run.

We performed a 3-fold cross-validation by training the decoder on two of the three averages for each location and then having the decoder predict the location on the basis of the remaining average for each location. We then repeated the process two more times with new decoders, changing which averages were used for training and which were used for testing. This process was then iterated 100 times, with different random subsets of trials used to create the averaged ERPs for each iteration. Completely new decoders were trained for every fold, iteration, time point, and participant.

For each fold, the decoder consisted of four separate SVMs, each of which was trained to distinguish between one location and the other three locations on the basis of the pattern of voltage across electrode sites. To test the decoder, the vector of voltages across electrode sites for a given location from the test average was passed to all four SVMs. The decoder predicted the location for that test average using MATLAB's predict() function, which minimized the average binary loss over the four SVMs. Decoding accuracy was computed as the proportion of correct predictions by the decoder across the three folds and 100 iterations. Because there were four locations, chance decoding accuracy was 25%.

We then averaged decoding accuracy across the cardinal and diagonal decoding runs, separately for oddballs and standards. This gave us a decoding accuracy for each participant at each time point for oddballs and for standards. To maximize statistical power, decoding accuracy was then averaged across an a priori time window that began at the start of the P3b measurement window (362 ms) and extended until the end of the epoch (1496 ms). For each participant, this gave us one decoding accuracy value for oddballs and another for standards.

Results

Behavioral results

Oddball task. We first examined the speed and accuracy for the buttonpress task, in which participants indicated whether a given target was an oddball or a standard. As is typical in oddball tasks, mean accuracy was lower for oddballs (80.5% ± 2.32%) than for standards (95.7% ± 0.512%), which was significant in a paired t test ($t(21) = 6.85$, $p < 0.0001$, $d_z = 1.46$). In addition, mean RTs were slower for oddballs (470.66 ± 19.48 ms) than for standards (346.04 ± 18.75 ms), which was also a significant difference ($t(21) = 14.59$, $p < 0.0001$, $d_z = 3.11$). Note that all statistical tests reported in this paper were two-tailed and used an alpha of .05. Effect sizes are quantified as d_z, which is the standard effect size metric corresponding to paired and one-sample t tests (Cohen J, 1988; Lakens, 2013). When means are given, we also provide the standard error of the mean (SEM).

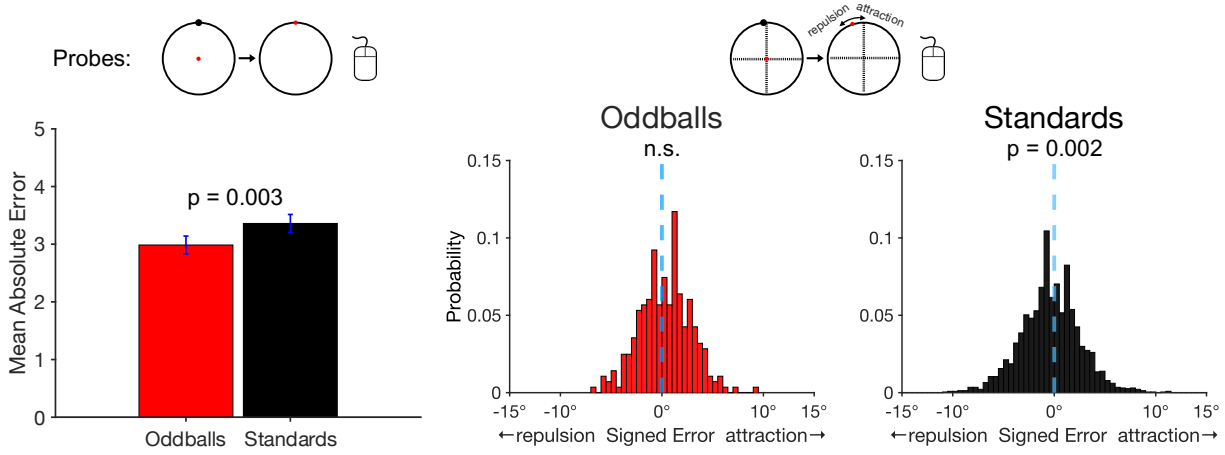


Figure 2. Probe trial results. A) Mean absolute error in the report of the exact location of the target disc. Error bars show ±1 SEM. B) Histograms of the single-trial bias values for target discs near a cardinal axis. For oddballs, the distribution of errors was fairly symmetrical around zero (minimal bias). For standards, the distribution is shifted toward negative values (repulsion away from the axis).

Probe task. We next examined accuracy on the probe trials, for which participants clicked on the remembered location of the target. Figure 2a shows that the mean absolute response error for the probe task was smaller for oddballs (2.99 ± 0.155°) than for standards (3.36 ± 0.156°),

which was a significant difference in a paired t test ($t(21) = 3.30$, $p = 0.003$, $d_z = 0.704$). This result indicates that working memory representations were more accurate for rare stimuli than for frequent stimuli.

We also asked whether working memory representations of rare stimuli are less subject to systematic biases. In particular, we took advantage of the fact that stimuli presented near a cardinal axis are typically remembered as being shifted away from that axis (Bae, 2022; Pratte et al., 2017; Wei & Stocker, 2015). We asked whether this bias would be reduced for oddballs relative to standards. Figure 2b shows histograms of the single-trial bias values, with negative values indicating a bias away from the nearby cardinal axis and positive values indicating a bias toward the axis. On standard trials, the distribution of values was shifted toward the negative side of zero, indicating the typical finding of repulsion away from the axis. We computed an average bias score for each participant, and we found that this bias score was significantly different from zero for the standards in a one-sample t test ($t(21) = 14.59$, $p = 0.002$, $d_z = 0.739$). On oddball trials, the distribution of bias values was more symmetrical around zero, and the average bias score was not significantly different from zero ($t(21) = 1.08$, $p = 0.292$, $d_z = 0.231$). The key finding was that the average bias score was significantly more negative for standards than for oddballs in a paired t test ($t(21) = 3.17$, $p = 0.005$, $d_z = 0.676$). This indicates that working memory representations were less biased for rare stimuli in addition to being more accurate.

P3b amplitude

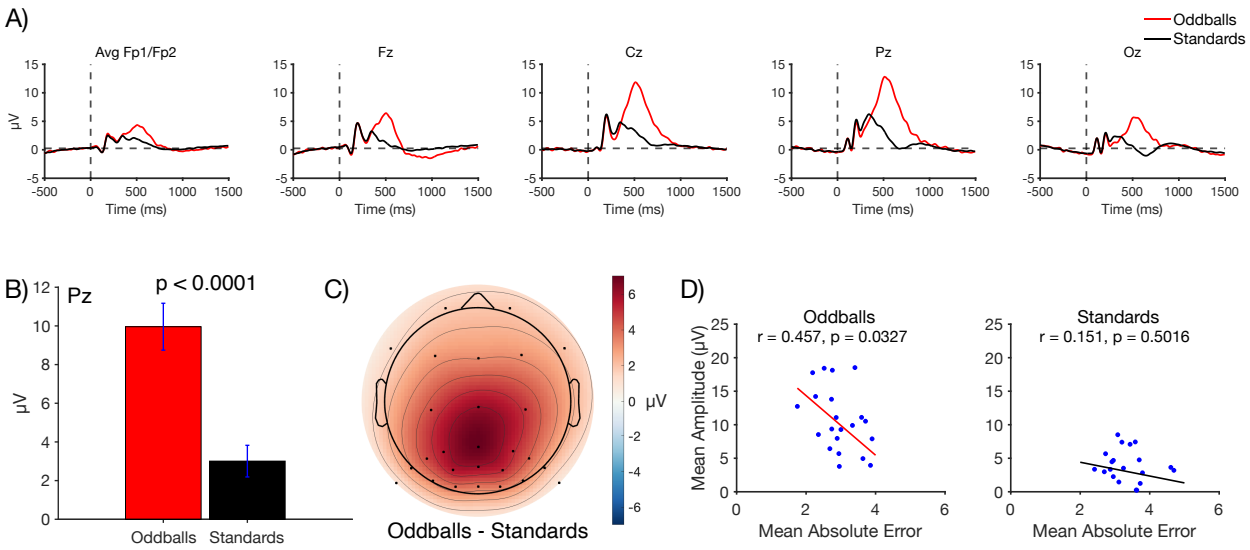


Figure 3. A) Grand average ERP waveforms for oddballs and standards at representative electrode sites. The P3b component peaked shortly after 500 ms and was largest at central and parietal electrode sites. B) Mean amplitude at site Pz measured across a 300-ms time window centered around the peak (362-662 ms). Error bars show ±1 SEM. C) Grand average scalp topography for the difference in amplitude between oddballs and standards in the P3b time window. D) Scatterplots showing the correlation between P3b amplitude (measured from 362-662 ms at the Pz electrode site) and behavioral accuracy on probe trials (mean absolute error). Each dot shows an individual participant, and the line is the best-fit regression line.

Figure 3a shows the grand average ERP waveforms for oddballs and standards at five representative electrode sites. As is typical, a large P3b was visible for oddballs, peaking at 512 ms at the Pz electrode site. As shown in Figure 3b, the mean voltage during the measurement window (362-662 ms) at Pz was significantly larger for oddballs than for standards in a paired t test ($t(21) = 8.10$, $p < 0.0001$, $d_z = 1.43$). The scalp distribution of the oddball-minus-standard difference wave during this window showed the typical midline centroparietal maximum (Figure 3c). Thus, although the oddball paradigm used in the present study was somewhat unusual, it yielded the typical pattern of a larger P3b for oddballs than for standards.

We also asked whether individuals with larger P3b amplitudes had more accurate working memory representations. Specifically, we examined the correlation between P3b

amplitude (mean voltage at Pz during the measurement window on all oddball or standard trials) and the absolute response error (on oddball or standard probe trials). The scatterplots are shown in Figure 3d. A statistically significant correlation was observed for oddballs ($r = 0.457$, $p = 0.033$), with a smaller error for participants with a larger P3b amplitude. A correlation in the same direction was observed for standards, but it was weak and not significant ($r = 0.151$, $p = 0.502$).

Together, these results show that oddball stimuli produce both a larger P3b and more accurate working memory representations, and that these effects are correlated with each other. Note, however, that this does not demonstrate a causal relationship between the P3b and the working memory enhancement. It is entirely plausible that both effects are results of a common underlying factor (e.g., an increase in attention triggered by the oddballs, which separately impacts P3b amplitude and working memory encoding).

Decoding Results

Our final analyses were designed to determine whether we could see evidence of enhanced working memory for oddballs in the brain activity measured between the onset of the P3b wave and the end of the trial. Toward that end, we attempted to decode the location of the stimulus from the ERP activity at each moment in time across the recording epoch. We collapsed across the two locations that were near a given axis (e.g., 44° and 46°), giving us four cardinal locations and four diagonal locations. We then decoded which of the four cardinal locations was presented when the cardinals were standards, which of the four cardinal locations was presented when the cardinals were oddballs, which of the four diagonal locations was presented when the diagonals were standards, and which of the four diagonal locations was presented when the diagonal were oddballs. We then collapsed across the diagonal and cardinal categories to obtain

overall decoding accuracy for oddballs and for standards. Because there were four locations within a category, chance decoding accuracy was 25%.

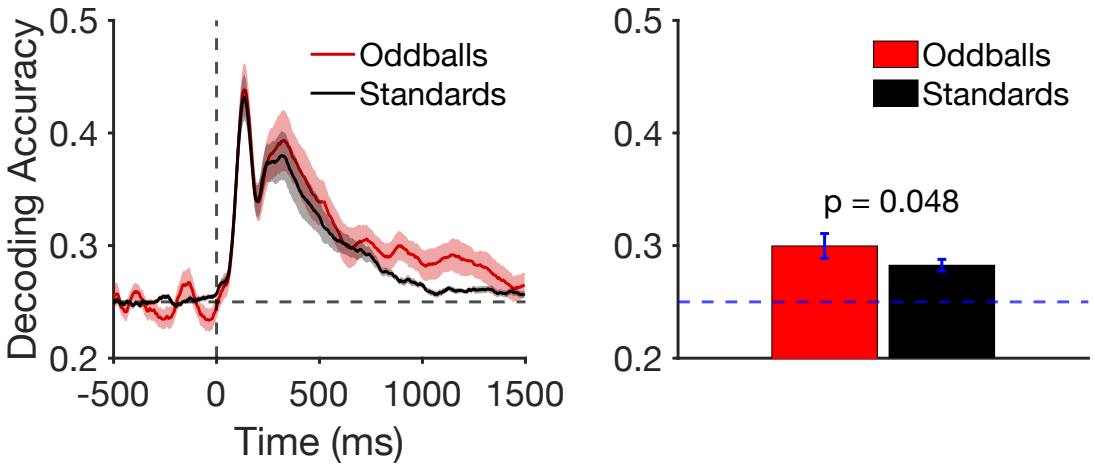


Figure 4. A) Timepoint-by-timepoint decoding of target location on oddball and standard trials. The dark lines show mean decoding accuracy across participants, and the shading shows ±1 SEM. B) Mean decoding accuracy between the start of the P3b measurement window (362 ms) and the end of the epoch (1500 ms). Error bars show ±1 SEM.

Figure 4a shows decoding accuracy at each individual time point. Location decoding accuracy was slightly greater for oddballs than for standards for much of the epoch, but most noticeably from approximately 800-1400 ms. To maximize statistical power, we averaged the decoding accuracy across an a priori time window that began at the start of the P3b measurement window (362 ms) and extended through the end of the epoch (1496 ms). As shown in Figure 4b, mean decoding accuracy was greater for oddballs than for standards, which was a significant difference in a paired t test ($t(21) = 2.10$, $p = 0.048$, $d_z = 0.425$). Thus, the representation of the target was modestly but significantly enhanced for oddballs relative to standards during the period of time following stimulus offset.

Discussion

The current study sought to provide direct evidence that working memory representations are enhanced for relatively rare task-relevant events compared to relatively frequent task-relevant events. Traditional oddball paradigms do not provide a sensitive assessment of working memory representations, but we were able to assess the accuracy of target representations behaviorally by including probe trials. We also assessed the neural representation of the target by decoding the target's location on the basis of the EEG activity following target offset.

As expected, oddballs elicited a much larger P3b wave than standards, with the prototypical P3b scalp distribution. This indicates that the modifications we made to the oddball paradigm did not disrupt its fundamental nature. The behavioral data from probe trials showed that participants maintained a more accurate and less biased representation of the target for oddballs than for standards. In addition, the EEG decoding results showed that the neural representation of the target was enhanced for oddballs relative to standards in the period following P3b onset. These findings are, to our knowledge, the first direct evidence that working memory is enhanced for rare, task-relevant stimulus categories.

The present results do not indicate the nature of the neural mechanism underlying enhanced working memory, but such an effect could potentially be a result of increased firing by neurons in the locus coeruleus, which would lead to a phasic increase in norepinephrine release throughout the cortex (Aston-Jones & Cohen, 2005; Krebs et al., 2018; Murphy et al., 2011; Nieuwenhuis et al., 2005). This is consistent with prior findings that norepinephrine is critical for signaling oddballs in humans and that norepinephrine depletion leads to working memory deficits in nonhuman primates (Arnsten, 2006; Arnsten & Goldman-Rakic, 1985; Brozoski et al., 1979; Cai et al., 1993; Franowicz et al., 2002; Strange & Dolan, 2007; Zhang et al., 2013).

The finding that working memory representations were enhanced for rare items—with greater accuracy for participants with larger P3b amplitudes—might be interpreted as evidence for the context updating hypothesis of the P3b wave, which was proposed over 40 years ago by Donchin (1981). Although Donchin did not use the term working memory, more recent researchers have interpreted the context updating hypothesis as stating that the P3b is related to working memory encoding (Kok, 2001; Polich, 2007). However, there is not much evidence to support the idea that the P3b specifically reflects the updating of working memory (Verleger, 2008). Indeed, using a reference-back task that was specifically designed to distinguish between updating and other processes, Rac-Lubashevsky and Kessler (2019) found no evidence that a P3b was triggered when working memory updating was required. Moreover, the P3b component in the present study was more than three times larger for oddballs than for standards, and yet behavioral accuracy on probe trials (a fairly direct measure of working memory) was only slightly greater for oddballs than for standards. Similarly, although decoding accuracy greater for oddballs than for standards, the difference in decoding accuracy was modest.

Thus, it is unlikely that the neural mechanisms that produce the P3b component are the same mechanisms that encode information into working memory. The correlation between P3b amplitude and working memory accuracy in the present study is more likely related to a shared impact of attentional allocation on P3b amplitude and working memory encoding. By analogy, pupil dilation is also enhanced for oddballs, correlated with working memory performance, and linked with locus coeruleus activity (Aminihajibashi et al., 2019; Eckstein et al., 2019; Gilzenrat et al., 2010; Murphy et al., 2011; Robison et al., 2023; Robison & Unsworth, 2019; Zokaei et al., 2019). However, this does mean that the mechanisms that produce pupil dilation are the same as the mechanisms that encode information into working memory. Instead, the P3b may be related

to models in which locus coeruleus activity provides a signal across the cerebral cortex that predictions about the environment have been strongly violated (Jordan, 2023). However, additional research would be needed to provide conclusive evidence of causal links between locus coeruleus activity and the P3b wave.

Behavioral responses in working memory tasks can be influenced by decision processes at the time of report in addition to encoding and maintenance processes. To provide additional information about the nature of the differences in working memory between rare and frequent stimuli, we decoded the location of the target being held in working memory using the scalp topography of the ERP signal during the interval immediately following the target, prior to the probe period. The finding of greater decoding accuracy for oddballs than for standards during this period suggests that rare stimuli are encoded and maintained more accurately than frequent stimuli (as opposed to an effect of rareness that is isolated to decision processes at the time of response). However, we cannot be certain that these ERP decoding effects reflect the same processes responsible for the improved behavioral performance for rare stimuli on probe trials. Indeed, the ERP decoding analyses were designed to distinguish among the four different locations within a given response category, which were separated by 90°, whereas the behavioral responses were typically within 5° of the true target location (see Figure 2A). The coarseness of the decoding process may partly explain why the differences in decoding accuracy between oddballs and standards were so small. We are not yet at the point where we can decode differences in stimulus location from scalp voltages with the same precision as the behavioral responses.

We would like to emphasize that the present study used a behavioral task to create the rare and frequent stimulus categories in the context of the behavioral task. These were not

categories that vary in their frequency or novelty outside of this task. Stimuli that are intrinsically novel elicit a very different pattern of neural activity, including the novelty P3a component (which has a much more frontal scalp distribution than the P3b component) and extensive BOLD activity in temporal and inferior frontal cortex (Strobel et al., 2008). It is not known how intrinsic novelty impacts working memory. Performance in visual working memory tasks is more accurate for familiar than for unfamiliar stimuli under some conditions (Chen et al., 2006; Jackson & Raymond, 2008; Ngiam et al., 2019), but this likely reflects the use of longer-term memory representations to aid in task performance. Click or tap here to enter text. Novelty can improve working memory under other conditions, particularly encoding processes (Mayer et al., 2011). Additional research will be necessary to fully understand how this kind of intrinsic novelty is related to the encoding and maintenance of information leading to enhancements in working memories.

Proactive interference provides a potential explanation for the finding of differences in working memory accuracy between the rare and frequent categories in the present study. In studies using verbal materials, working memory accuracy is reduced when similar information is presented across many trials (Baddeley, 1986; Keppel & Underwood, 1962). In the present study, locations in the frequent category were, by definition, presented more frequently than the locations in the rare category, which could have led to proactive interference. However, this kind of interference is not typically seen in visual working memory tasks using brief retention intervals like those in the current study (Lin & Luck, 2012; Oberauer et al., 2017). Future research would be needed to provide a conclusive test of this explanation.

Data and Code Availability

All data will be available at: doi.org/10.17605/OSF.IO/HV7JU, upon publication.

Author Contributions

Carlos Daniel Carrasco: conceptualization, analyses, writing; Aaron Matthew Simmons: software development, writing; John E. Kiat: analyses, writing; Steven J. Luck: conceptualization, writing, development of methods, funding acquisition. Editing/Review was performed by all authors prior to submission of the final version of the manuscript.

Acknowledgements

We thank the Luck lab for all the help and support always.

Funding

This study was supported by grants R01MH087450, R01EY033329, and R01MH076226 from the National Institute of Mental Health.

References

Aminihajibashi, S., Hagen, T., Foldal, M. D., Laeng, B., & Espeseth, T. (2019). Individual differences in resting-state pupil size: Evidence for association between working memory capacity and pupil size variability. International Journal of Psychophysiology, 140, 1–7. https://doi.org/10.1016/j.ijpsycho.2019.03.007

Arnsten, A. F. T. (2006). Fundamentals of Attention-Deficit/Hyperactivity Disorder: Circuits and Pathways. In J Clin Psychiatry (Vol. 67, Issue 8).

Arnsten, A. F. T., & Goldman-Rakic, P. S. (1985). α2-Adrenergic Mechanisms in Prefrontal Cortex Associated with Cognitive Decline in Aged Nonhuman Primates. In New Series (Vol. 13, Issue 4731).

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. In Annual Review of Neuroscience (Vol. 28, pp. 403–450). https://doi.org/10.1146/annurev.neuro.28.061604.135709

Baddeley, A. (1986). Working memory. Oxford: Oxford University Press, Clarendon Press.

Bae, G. Y. (2022). Breaking the cardinal rule: The impact of interitem interaction and attentional priority on the cardinal biases in orientation working memory. Attention, Perception, and Psychophysics, 84(7), 2186–2194. https://doi.org/10.3758/s13414-021-02374-2

Bae, G. Y., & Luck, S. J. (2019). Decoding motion direction using the topography of sustained ERPs and alpha oscillations. NeuroImage, 184, 242–255. https://doi.org/10.1016/j.neuroimage.2018.09.029

Bae, G. Y. (2021). The Time Course of Face Representations during Perception and Working Memory Maintenance. Cerebral Cortex Communications, 2(1), 1–12. https://doi.org/10.1093/texcom/tgaa093

Bae, G. Y., & Luck, S. J. (2018). Dissociable Decoding of Spatial Attention and Working Memory from EEG Oscillations and Sustained Potentials. The Journal of Neuroscience, 38(2), 2860–17. https://doi.org/10.1523/JNEUROSCI.2860-17.2017

Bledowski, C., Prvulovic, D., Goebel, R., Zanella, F. E., & Linden, D. E. J. (2004). Attentional systems in target and distractor processing: A combined ERP and fMRI study. NeuroImage, 22(2), 530–540. https://doi.org/10.1016/j.neuroimage.2003.12.034

Brozoski, T. J., Brown, R. M., Rosvold, H. E., & Goldman, P. S. (1979). Cognitive Deficit Caused by Regional Depletion of Dopamine in Prefrontal Cortex of Rhesus Monkey. In New Series (Vol. 205, Issue 4409).

Cai, J. X., Ma, Y.-Y., & Hu, X.-T. (1993). Reserpine impairs spatial working memory performance in monkeys: reversal by the az-adrenergic agonist clonidine. In Brain Research (Vol. 614).

Chen, D., Yee Eng, H., & Jiang, Y. (2006). Visual working memory for trained and novel polygons. Visual Cognition, 14(1), 37–54. https://doi.org/10.1080/13506280544000282

Clark, V. P., Fannon, S., Lai, S., Benson, R., & Bauer, L. (2000). Responses to Rare Visual Target and Distractor Stimuli Using Event-Related fMRI.

Cohen J. (1988). Statistical Power Analysis for the Behavioral Sciences Second Edition.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. In Journal of Neuroscience Methods (Vol. 134). http://www.sccn.ucsd.edu/eeglab/

Dietterich, T. G., & Bakiri, G. (1995). Solving Multiclass Learning Problems via Error-Correcting Output Codes. In Journal of Artiicial Intelligence Research (Vol. 2).

Donchin, E. (1981). Surprise!...Surprise? Psychophysiology, 18(5), 493–513. https://doi.org/10.1111/j.1469-8986.1981.tb01815.x

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? Behavioral and Brain Sciences, 11(3), 357–374. https://doi.org/10.1017/S0140525X00058027

Eckstein, M. K., Starr, A., & Bunge, S. A. (2019). How the inference of hierarchical rules unfolds over time. Cognition, 185, 151–162. https://doi.org/10.1016/j.cognition.2019.01.009

Franowicz, J. S., Kessler, L. E., Dailey Borja, C. M., Kobilka, B. K., Limbird, L. E., & Arnsten, A. F. T. (2002). Mutation of the 2A-Adrenoceptor Impairs Working Memory Performance and Annuls Cognitive Enhancement by Guanfacine.

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. Cognitive, Affective and Behavioral Neuroscience, 10(2), 252–269. https://doi.org/10.3758/CABN.10.2.252

Jackson, M. C., & Raymond, J. E. (2008). Familiarity Enhances Visual Working Memory for Faces. Journal of Experimental Psychology: Human Perception and Performance, 34(3), 556–568. https://doi.org/10.1037/0096-1523.34.3.556

Johnson, R. (1988). SCALP-RECORDED P300 ACTIVITY IN PATIENTS FOLLOWING UNILATERAL TEMPORAL LOBECTOMY. In Brain (Vol. 111). https://academic.oup.com/brain/article/111/6/1517/297108

Jordan, R. (2023). The locus coeruleus as a global model failure system. In Trends in Neurosciences. Elsevier Ltd. https://doi.org/10.1016/j.tins.2023.11.006

Katayama, J., & Polich, J. (1998). Stimulus context determines P3a and P3b. Psychophysiology, 35(1), 23–33. https://doi.org/10.1111/1469-8986.3510023

Keppel, G., & Underwood, B. J. (1962). Proactive Inhibition in Short-Term Retention of Single Items 1.

Kim, H. (2014). Involvement of the dorsal and ventral attention networks in oddball stimulus processing: A meta-analysis. Human Brain Mapping, 35(5), 2265–2284. https://doi.org/10.1002/hbm.22326

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. Psychophysiology, 38(3), 557–577. https://doi.org/10.1017/S0048577201990559

Krebs, R. M., Park, H. R. P., Bombeke, K., & Boehler, C. N. (2018). Modulation of locus coeruleus activity by novel oddball stimuli. Brain Imaging and Behavior, 12(2), 577–584. https://doi.org/10.1007/s11682-017-9700-4

Kutas, M., Mccarthy, G., & Donchin, E. (1977). Augmenting Mental Chronometry: The P300 as a Measure of Stimulus Evaluation Time. In Source: Science (Vol. 197, Issue 4305).

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4(NOV). https://doi.org/10.3389/fpsyg.2013.00863

Lin, P. H., & Luck, S. J. (2012). Proactive interference does not meaningfully distort visual working memory capacity estimates in the canonical change detection task. Frontiers in Psychology, 3(FEB). https://doi.org/10.3389/fpsyg.2012.00042

Linden, D. E. J., Prvulovic, D., Formisano, E., Völlinger, M., Zanella, F. E., Goebel, R., & Dierks, T. (1999). The functional neuroanatomy of target detection: an fMRI study of visual and auditory oddball tasks. Cerebral Cortex, 9(8), 815–823.

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. Frontiers in Human Neuroscience, 8(1 APR), 1–14. https://doi.org/10.3389/fnhum.2014.00213

Luck, S. J. (2014). An Introduction To The Event-Related Potential Technique (2nd Editio). MIT Press Journals.

Luck, S. J. (2023). ERPLAB Decoding Tutorial. https://github.com/ucdavis/erplab/wiki/ERPLAB-Decoding-Tutorial

Mayer, J. S., Kim, J., & Park, S. (2011). Enhancing visual working memory encoding: The role of target novelty. Visual Cognition, 19(7), 863–885. https://doi.org/10.1080/13506285.2011.594459

Mecklinger, A., & Ullsperger, P. (1993). P3 varies with stimulus categorization rather than probability. In Electroencephalography and clinical Neurophysiology (Vol. 86).

Murphy, P. R., Robertson, I. H., Balsters, J. H., & O'connell, R. G. (2011). Pupillometry and P3 index the locus coeruleus-noradrenergic arousal function in humans. Psychophysiology, 48(11), 1532–1543. https://doi.org/10.1111/j.1469-8986.2011.01226.x

Ngiam, W. X. Q., Khaw, K. L. C., Holcombe, A. O., & Goodbourn, P. T. (2019). Visual working memory for letters varies with familiarity but not complexity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 45(10), 1761.

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus-norepinephrine system. In Psychological Bulletin (Vol. 131, Issue 4, pp. 510–532). https://doi.org/10.1037/0033-2909.131.4.510

Nieuwenhuis, S., De Geus, E. J., & Aston-Jones, G. (2011). The anatomical and functional relationship between the P3 and autonomic components of the orienting response. In Psychophysiology (Vol. 48, Issue 2, pp. 162–175). Blackwell Publishing Inc. https://doi.org/10.1111/j.1469-8986.2010.01057.x

Oberauer, K., Awh, E., & Sutterer, D. W. (2017). The role of long-term memory in a test of visual working memory: Proactive facilitation but no proactive interference. Journal of Experimental Psychology: Learning Memory and Cognition, 43(1), 1–22. https://doi.org/10.1037/xlm0000302

Paller, K. A., Mccarthy, G., Roessler, E., Allison, T., & Wood, C. C. (1992). Potentials evoked in human and monkey medial temporal lobe during auditory and visual oddball paradigms. In Electroencephalography and clinical Neurophysiology (Vol. 84).

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. Behavior Research Methods, 51(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Polich, J. (1986). A'ITENTION, PROBABILITY, AND TASK DEMANDS AS DETERMINANTS OF P300 LATENCY FROM AUDITORY STIMULI t. In Electroencephalograph)' and clinical Neurophysiology (Vol. 63).

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. Clinical Neurophysiology, 118(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Polich, J. (2012). Neuropsychology of P300. The Oxford Handbook of Event-Related Potential Components, 159–188.

Pratte, M. S., Park, Y. E., Rademaker, R. L., & Tong, F. (2017). Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. Journal of Experimental Psychology: Human Perception and Performance, 43(1), 6–17. https://doi.org/10.1037/xhp0000302

Rac-Lubashevsky, R., & Kessler, Y. (2019). Revisiting the relationship between the P3b and working memory updating. Biological Psychology, 148. https://doi.org/10.1016/j.biopsycho.2019.107769

Robison, M. K., Ralph, K. J., Gondoli, D. M., Torres, A., Campbell, S., Brewer, G. A., & Gibson, B. S. (2023). Testing locus coeruleus-norepinephrine accounts of working memory, attention control, and fluid intelligence. Cognitive, Affective and Behavioral Neuroscience, 23(4), 1014–1058. https://doi.org/10.3758/s13415-023-01096-2

Robison, M. K., & Unsworth, N. (2019). Pupillometry tracks fluctuations in working memory performance. Attention, Perception, and Psychophysics, 81(2), 407–419. https://doi.org/10.3758/s13414-018-1618-4

Soltani, M., & Knight, R. T. (2000). Neural Origins of the P300. In Critical Reviews in Neurobiology (Vol. 14).

Strange, B. A., & Dolan, R. J. (2007). β-adrenergic modulation of oddball responses in humans. Behavioral and Brain Functions, 3. https://doi.org/10.1186/1744-9081-3-29

Strobel, A., Debener, S., Sorger, B., Peters, J. C., Kranczioch, C., Hoechstetter, K., Engel, A. K., Brocke, B., & Goebel, R. (2008). Novelty and target processing during an auditory novelty oddball: A simultaneous event-related potential and functional magnetic resonance imaging study. NeuroImage, 40(2), 869–883. https://doi.org/10.1016/j.neuroimage.2007.10.065

Verleger, R. (2008). P3b: Towards some decision about memory. In Clinical Neurophysiology (Vol. 119, Issue 4, pp. 968–970). https://doi.org/10.1016/j.clinph.2007.11.175

Wei, X. X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. Nature Neuroscience, 18(10), 1509–1517. https://doi.org/10.1038/nn.4105

Zhang, Z., Cordeiro Matos, S., Jego, S., Adamantidis, A., & Séguéla, P. (2013). Norepinephrine Drives Persistent Activity in Prefrontal Cortex via Synergistic α1 and α2 Adrenoceptors. PLoS ONE, 8(6). https://doi.org/10.1371/journal.pone.0066122

Zokaei, N., Board, A. G., Manohar, S. G., & Nobre, A. C. (2019). Modulation of the pupillary response by the content of visual working memory. Proceedings of the National Academy of Sciences of the United States of America, 116(45), 22802–22810. https://doi.org/10.1073/pnas.1909959116

Chapter 4: Reinstatement of Working Memory Representations Facilitates Credit Assignment

Carlos Daniel Carrasco, Yanchang Zhang, Steven J. Luck & Erie D. Boorman

Center for Mind & Brain, University of California, Davis

*This work is currently being prepped for submission:

A preprint will be available upon submission.

CC-By Attribution 4.0 International, https://creativecommons.org/licenses/by/4.0/.

Abstract

Neural representations of credit assignment processes have been seen in an array of brain regions, predominantly frontal, parietal, and temporal areas related to value computations and updating. Working memory processes in turn are used for the maintenance, manipulation, and updating of information for goal driven behaviors implicating them in credit assignment processes. What is not known is if the initial neural representations of working memories related to choice and counterfactual feedback items are reactivated in working memory to facilitate credit assignment after an outcome occurs. To address this, we designed a 2-armed bandit task with independent slowly drifting probabilities of reward that encouraged model-based credit assignment. We paired this with a localizer task and scalp EEG recordings to decode the information content during sensory and feedback epochs. We found behavioral evidence that participants integrated information about the slowly drifting probabilities of reward and neural evidence that information related to choices and counterfactuals was reactivated at the time of feedback where participants were informed of outcomes.

Significance Statement

Contingency learning is often associated with frontal, temporal and parietal areas, but the nature of the information content is often speculated to be a conglomerate of higher order value computations that update iteratively through the process of learning. In this study we provide direct evidence that sensory information content is reinstated in neural signals at outcome epochs to facilitate credit assignment as participants learn the consequences of their choices in a novel bandit task. By modeling the task via the use of a Bayesian learner, we ensured that the neural learning signals we decoded were not being driven by biases in participants favoring one stimulus over the other.

Introduction

Contingency learning—through which beings form associations between stimuli and outcomes—is supported by credit assignment systems that have been formalized in various ways. One type of credit assignment system is habitual or model-free, in which the most recent actions are assigned credit for a given outcome. This is slow, computationally inexpensive, and supplemented by weighted exponentially decaying neural memory representations known as eligibility traces (Dayan & Berridge, 2014; Gerstner et al., 2018; He et al., 2015; Pan et al., 2005). The other type of credit assignment system is often termed as being model based, in which a more complex model of the environment is used to link actions with outcomes. This system is fast, computationally expensive, tied to a model, and is theorized to involve a causal reinstatement of past choices when an outcome occurs (Dayan & Berridge, 2014; Dayan & Niv, 2008; Moran et al., 2019). Many studies have implicated mesolimbic reward system circuitry inputting into the cortico-basal ganglia-thalamic network in credit assignment (Deserno et al., 2021; Moran et al., 2021; Noonan et al., 2017; Richards & Lillicrap, 2019; Witkowski et al., 2022). Additionally, working memory has been suggested to be the primary model-based system used in credit assignment, because it has the ability to flexibly update rulesets and allows for dynamic switching in a model-based ways (Collins & Cockburn, 2020; Collins & Frank, 2012, 2018; Yoo & Collins, 2022). However, evidence is lacking for the reinstatement of information in working memory during credit assignment. The present study sought to determine if using scalp EEG paired with multivariate pattern analysis, we could decode the information content related to a given stimulus at feedback when the stimuli were no longer on the screen and if said neural representations contained any detectable information related to the sensory features of a given item.

We designed a 2-armed bandit task with two independent arms that had slowly drifting reward schedules with randomly generated magnitudes from 10-100 points. On each trial, a face and a house were shown (corresponding to the two arms of the bandit), with the number of points that could potentially be won by choosing that option (Figure 1A). The participants then chose either the face or house. After a long delay, feedback screens indicated whether the chosen stimulus generated a reward on that trial (green versus red circle) and whether the nonchosen stimulus would have generated a reward if it had been chosen (green versus red square), along with the number of points associated with the reward. In the example shown in Figure 1A, the face had a potential value of 20 points and the house had a potential value of 35 points. The participant chose the house, and the house was actually rewarded with 35 points (indicated by the green circle with 35 in the middle). If the participant had chosen the face, it would not have been rewarded on that trial (indicated by the red square). Thus, participants received feedback about both the chosen stimulus and the nonchosen (counterfactual) stimulus. This design made it possible to determine, from the EEG signals, whether the chosen stimulus was maintained in working memory during the delay period prior to the feedback and whether a given stimulus was reinstated in working memory following each feedback stimulus to link it with the outcome. Specifically, we asked whether we could decode which stimulus had been chosen, both during the delay period and following each feedback stimulus. We predicted that participants would reinstate rather than maintain the neural representations of a given choice and the counterfactual at feedback. Additionally, because we included a sensory "localizer" task at the beginning of the session, we were able to ask whether the reinstatement consisted of a reactivation of the sensory features of the stimulus. We tested this by training the decoder on the localizer data and then testing with the data from the main task.

EEG data can be analyzed in terms of the time-locked voltage at a given time point or the magnitude of an oscillation around that time point. Both time-locked voltages and alpha-band (10 Hz) oscillations can be decoded (Bae & Luck, 2019; Bae & Luck, 2018; Nadra et al., 2023; Noah et al., 2020). We therefore analyzed both the time-locked voltages and the alpha band power.

Methods

Participants

Twenty-eight human participants, (21 females, 7 males) with a mean age of 21.3 (SD = 2.03), from the UC Davis community who provided informed consent completed the study. All participants were neurotypical and had normal or corrected to normal vision with no history of neurological conditions. Twenty-seven participants were right-handed, and one was left-handed. Twenty-one participants were included in the study after exclusions for artifacts in EEG data and learning criteria. Participants were excluded if their EEG data had < 25% trials rejected (N = 3) or they were deemed to not be optimally learning if they did not pick the optimal choice > 80% of the time in the study (N = 4). Monetary compensation was given at a rate of $15/hr. The study was approved by the University of California Davis, Institutional Review Board.

Stimuli and Task

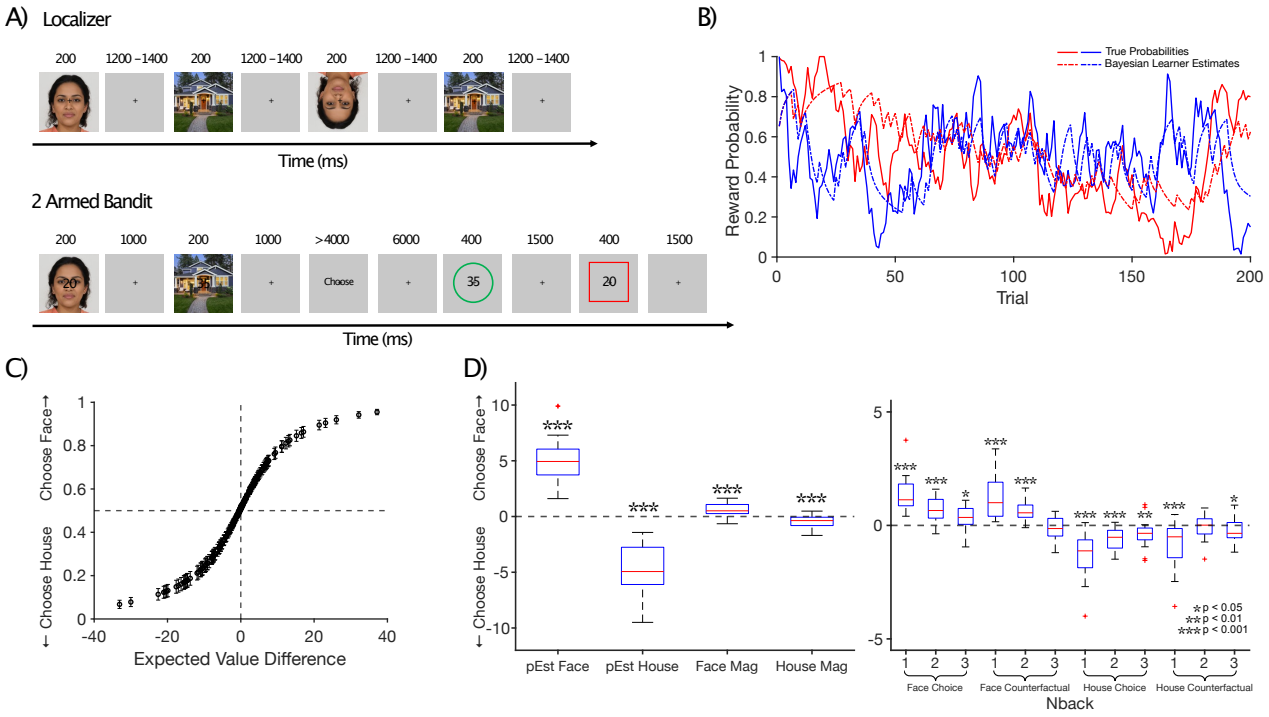


Figure 1. Overview of study and behavior. A) Participants completed a localizer task where we collected sensory EEG signals elicited by both stimuli. To ensure subject engagement, subjects were required to detect occasional upside-down stimuli. In the 2-armed bandit task, participants did their best to track the slowly drifting independent probabilities of each stimulus to earn the most points possible on a trial-by-trial basis by choosing the item they believed would lead to more favorable reward on each trial based on the feedback from the previous trials. B) The slowly drifting probabilities of reward for each stimulus, and the learner optimal Bayesian learner's estimates of these probabilities. C) Group level choice probabilities that are predicted by the expected value differences via a sigmoid fit. The point of indifference between the chosen stimulus and the other option is y = 0.5. D) Left: GLM with logit link function indicating that participants' choices were influenced by the probability estimates and magnitudes associated with each stimulus independently. Right: A GLM with the same logit link function regressing the individual feedback against a participant's individual choices indicating that participants were integrating information about a given outcome up to 3 trials back whether it was the chosen option, or the counterfactual.

Localizer Task

Participants were presented with stimuli using Psychopy (Peirce et al., 2019) on an LCD monitor (Dell U2412M) with a gray background (27.7 cd/m^2) at a viewing distance of 100 cm. A continually visible fixation cross (Thaler et al., 2013) was kept at the center of the screen, where

participants fixated throughout the task. Figure 1A (top) shows the localizer paradigm, which consisted of the presentation of the face and house stimuli that would be used in the main task. A stimulus was presented upright (p = .8) or inverted (p = .2) for 200 ms on each trial, followed by a 1200 – 1400 ms blank period. During the blank period, the participant pressed one of two buttons for either stimulus using the arrow keys on a keyboard with their right hand (counterbalanced across participants) to indicate whether the stimulus had been upright or inverted. The face was generated using NVIDIAs face photo generator (NVIDIA, 2019) and we used a stock image of a house. We chose face and house stimuli given the known evidence of the distinct representations of faces and houses (O'toole et al., 2005). A total of 360 trials were presented (128 upright faces and houses; 32 inverted faces and houses)

2-Armed Bandit Task

Stimuli were presented using Psychopy (Peirce et al., 2019) on an LCD monitor (Dell U2412M) with a gray background (27.7 cd/m^2) at a viewing distance of 100 cm. As illustrated in Figure 1A, participants played a 2-armed bandit task with each arm having an independent drifting probability. Participants were instructed to choose the stimulus (arm) that would lead to the greatest reward. They were presented with a sequence of a 200-ms face and a 200-ms house (identical to those used in the localizer task) in a randomized order, separated by a 1000-ms ISI. A number of points was overlayed on each stimulus. After an additional 1000-ms ISI, the word 'Choose' appeared, and participants made a choice between the two stimuli by pressing the left and right arrow keys on a keyboard, The choice display had an open response window of up to 4000 ms otherwise the participants would see a screen that they took too long to decide. After a 6000-ms delay, they received a 400-ms feedback display containing a circle that was green for a correct response or red for an incorrect response. The circle contained the number of points that

they won or failed to win for this choice. After a 1.5 s ISI, they received a feedback square related to the other option, known as the counterfactual. This square was green to indicate that choosing that item would have been rewarded or red to indicate that it would not have been rewarded, along with the number of points that would have been won if that item had been chosen and rewarded. Note that each choice had an independent reward probability, such that neither, either, or both could lead to a reward if chosen on a given trial. The reward magnitudes were randomly generated for each stimulus on every trial and varied from 10 – 100 points. The probability of reward for a given choice drifted gradually across the session (as described below), which incentivized the participants to continually remain engaged and tracking the reward probabilities.

Trials were broken up into 5 blocks of 40 trials, resulting in 200 total trials. The points were summed for each individual block, and the sum for one of the blocks was randomly selected at the end of the session and transformed into a bonus cash payout (with an amount ranging from $7 – 15, depending on the number of points accrued in the chosen block).

To create reward schedules that would produce balanced choices, we generated 1000 random reward schedules and examined the choices made for each schedule by a Bayesian genetic learning algorithm; for details on the learner, see computational models section (Behrens et al., 2007; Boorman et al., 2009, 2011). The goal was to find a schedule in which the two outcomes were chosen approximately equally, and a given outcome was repeated no more than 3 times in a row less than 25% of the time. In the best reward schedule, the Bayesian learner learned a correlation between stimuli < 0.2 and learned a correlation of their independent magnitudes of < 0.2. The reward probabilities of this schedule are shown in Figure 1C, which

also shows that the probability estimates made by the model tracked the true reward probabilities well. This schedule was therefore used in the main experiment.

Participants first played a practice block of the bandit task, followed by taking a 4-question quiz to ensure that they understood the rules of the task. Only participants who made optimal choices in the practice block (> 80% accuracy) and completed the quiz correctly were allowed to complete the study.

Behavioral Analyses

To confirm that participants behavioral choices were balanced, we computed the mean number of times each stimulus was chosen. We then looked at the likelihood of a participant choosing a given option and computed the expected value difference between the two options, which we calculated by multiplying the magnitudes of each option and the probability estimates of the Bayesian learner on a given trial followed by computing the difference between the two options. Using these two metrics, we fit a generalized linear model (GLM) with a logit link function to compute the beta weight for a given participant that predicted their choice on an individual trial. Finally, we used these weights to project back the likelihood that a participant one chose option over the other option as a function of the expected value difference on a given trial; for details see computational models section.

We computed two additional logistic GLMs to assess how participants were integrating information about the stimuli and the influence of the past on the current trial. The first model regressed the probability estimates of either stimulus from the Bayesian learner, and their respective magnitudes against the individual participant's choices. The second model regressed the participant's choices against the past, up to three trials back, separately for each stimulus and their respective feedback. This meant that we could see the influence of both the chosen

outcomes and the counterfactuals separately; for details, see computational models section. Significance was assessed for these models by using a *t* test against chance, which in this case was 0 (i.e., no influence of the predictor on choosing either of the options).

EEG Recording and Preprocessing

Continuous EEG recordings were obtained using a Brain Products actiCHamp recording system (Brain Products) with 32 electrodes following the standard 10/20 system. Scalp electrodes included: FP1, Fz, F3, F7, Cz, C3, Pz, P3, P5, P7, P9, PO7, PO3, O1, POz, Oz, FP2, F4, F8, C4, P4, P6, P8, P10, PO4, PO8, O2. Five electrodes were placed at non-scalp locations: Three were electrooculogram electrodes capturing eyeblinks and eye movements (HEOG R, HEOG L, VEOG L). The remaining two electrodes were placed at the mastoids to serve as references (Mastoid R, Mastoid L). All electrode impedances were kept < 50 KΩ. EEG signals were recorded single-ended and referenced offline. Online EEG filtering was performed with a cascaded integrator-comb antialiasing filter (half-power cutoff at 130 Hz), and the EEG was digitized at 500 Hz. Preprocessing consisted of first shifting the triggers to accommodate for the delay between the triggers and the visual presentation (56 ms, as measured with a photodiode). All signals were then resampled at 250 Hz. The DC offset was removed, and the signals were high-pass filtered (noncausal Butterworth impulse response function, half-amplitude cutoff at 0.1 Hz, 12 dB/oct roll-off). Time segments between block were deleted, and the data were referenced to P9. Independent component analysis (ICA) was then performed to remove eyeblink and eye movement components. The data that were used for computing the ICA weights were filtered more aggressively with a bandpass filter (noncausal Butterworth impulse response function, half-amplitude cutoffs at 1 and 30 Hz, 48 dB/oct roll-off) and resampled at 100 Hz. The criterion for excluding an ICA component was consistency between the shape, timing and spatial

location of the component compared to HEOG and VEOG signals. ICA weights were transferred back to the 250 Hz data, and the blink and saccade components were then removed. ICA-corrected data were then re-referenced to the average of the left and right mastoid electrodes and segmented from -500 to 1496 ms relative to stimulus onset. For the delay period following the choice, the data were downsampled to 50 Hz and the epochs went from -500 to 5580 ms. Finally, artifact detection was applied to the segmented data to identify and flag any trials that had high voltage artifacts, or trials where the participant made eyeblinks at stimulus onset, using a moving window peak-to-peak function with a 200 ms moving window with a 100 ms window step on bipolar channels. Bipolar channels were created by subtracting the Fp2 electrode from the VEOG electrode and the two HEOG electrodes from one another. We rejected any participant who had > 25% artifact trials in a given bin. All preprocessing steps were conducted in MATLAB using the EEGLAB and ERPLAB toolboxes (Delorme & Makeig, 2004; Lopez-Calderon & Luck, 2014; MathWorks, 2023).

Support Vector Machine (SVM) Decoding Analyses

We used the pattern of voltage over the scalp electrodes at a given latency to decode whether the choice on a given trial was the face or the house. Prior to decoding, the EEG epochs were sorted according to which stimulus was chosen on a given trial, and we divided the sorted EEG data into two streams as in (Bae & Luck, 2018). One stream was low-pass filtered at 20 Hz and was used for time-locked decoding. The other stream was filtered from 8-12 Hz and passed through the Hilbert transform and squared to provide the alpha-band EEG magnitude (Bae & Luck, 2018). All reference electrodes, ocular electrodes, and bipolar channels were left out of the analyses. We performed SVM decoding using scripts calling decoding routines in ERPLAB Toolbox, which rely on MATLAB's fitcsvm() function with a linear kernel. Decoding was

performed separately at each time point. We ran all analyses on both the raw time-locked voltages and the instantaneous alpha band power.

Localizer Decoding

For a given participant, we first floored the number of trials to the stimulus bin with the lower number of trials. On average, participants had 114.24 ± 2.3 trials per stimulus. We then split those trials into 9 different random sets to create 9 different averaged waveforms for each stimulus. This resulted in 12.24 ± .025 trials going into the creation of each average waveform, conforming to the general range of average trials [10-20] that has been seen to work well in previous studies (Bae & Luck, 2018; Carrasco et al., 2024; Grootswagers et al., 2017). Once we had these waveforms, we trained and tested independent SVM classifiers to differentiate between when the face was presented and when the house was presented with a leave one out approach, meaning that we used 8 of the averaged waveforms for each class to train the classifiers and then tested on the one that had been left out from each class. We repeated this process until each of the waveforms was used as both a training and testing set. After this, we iterated this process 100 times, each time with a different random assignment of trials to the 9 averages. We then computed the decoding accuracy by averaging the accuracy over iterations for a given participant at a given time point. To compare the decoding accuracy of the group of participants to chance, we used a nonparametric cluster-based permutation test that found clusters of timepoints that were significantly different from a null distribution (Bae & Luck, 2019). Details on the statistical analysis of the decoding are described further in the decoding statistical tests section.

Bandit Task Decoding

In the bandit task, we primarily focused on decoding the difference in neural signals immediately after participants chose the stimulus and following the feedback screens. To see if

there was a difference between the neural representations held in mind during the long delay, we binned the 6 second epoch according to when participants either chose the face or the house. For each individual participant we then floored the number of trials to the bin with the lower number of trials and created averaged ERP waveforms to train and test on at each timepoint independently. In the case of choice decoding, we ended up with a mean of 89.5 ± 1.67 trials per stimulus and made 6 average ERP waveform sets which resulted in 14.5 ± 0.281 trials going into each waveform. We iterated this process 100 times to improve our decoding resolution. Since our epoch was so long compared to our other decoding analyses, we downsampled the data to 50 Hz (1 timepoint every 20 ms).

We next decoded the signals obtained following the feedback screens. This decoding was performed from -500 to 1496 ms in 4 ms) increments, with time 0 being the onset of the feedback screen. We ended up with a total of 40.95 ± 7.36 trials for win outcomes per stimulus, 34.43 ± 8.23 trials for loss outcomes per stimulus, 37 ± 4.46 counterfactual wins per stimulus, and 43.48 ± 6.06 counterfactual losses per stimulus. Due to the number of trials, we chose to do 3-fold cross validation which results in 13.65 ± 2.45 trials for win outcomes, 11.48 ± 2.74 trials for loss outcomes, 12.33 ± 1.49 trials for counterfactual wins, and 14.49 ± 2.02 trials for counterfactual losses going into the creation of each averaged waveform. Just as in the other decoding analyses, we left one average out of training and used it for testing, repeating the process till all waveforms had been used for testing. In the case of all decoding analyses a non-parametric cluster-based permutation test was used to assess significance.

Temporal Generalization

To assess if the representations being decoded at feedback epochs were made up of the same sensory features as the representations decoded using the localizer data, we performed a

cross-task decoding analysis. Specifically, we took a given participant's localizer data and used it for training, and then tested on that participant's bandit task data following the feedback displays. We could not assume that the timing of a given neural pattern following feedback would be the same as the timing following stimulus onset in the localizer task. Consequently, we performed temporal generalization decoding where we trained on a given localizer timepoint and tested across all timepoints following feedback onset in the bandit task. Due to the large number of decoders required for temporal generalization, we subsampled our data to 50 Hz (1 timepoint every 20 ms). We floored the number of trials for training and testing to the minimum number of trials in either of the feedback conditions being decoded. This means that the same number of trials for training and testing were used as in the bandit task feedback decoding analyses. For brevity this meant that there were: 13.65 ± 2.45 trials for win outcomes, 11.48 ± 2.74 trials for loss outcomes, 12.33 ± 1.49 trials for counterfactual wins, and 14.49 ± 2.02 trials for counterfactual losses going into the creation of each averaged waveform except, that now training waveforms were generated using the sensory voltages from the localizer task, and the testing voltages still remained the feedback screens of interest. An important step in both voltage and alpha decoding was to demean the datasets so that the decoding could ignore differences in other sources of neural activity during the training data and testing data. This resulted in four heatmap representations where each row represents a given training timepoint and each column all task testing timepoints. The intensity of the color indicates the probability of decoding a given representation at a given period of time. As in previous decoding analyses, permutation testing was used to test for significance.

Decoding Statistical Tests

For all decoding statistical tests, we used non-parametric, cluster-based permutations, an approach that is described in detail in (Bae & Luck, 2019). We took the decoding accuracy and compared it to chance via independent one-sample *t* tests for each timepoint. Following this, we then found clusters of adjacent timepoints that were significant at $p < 0.05$ and summed their *t* scores to produce a cluster-level *t* mass. Each of these cluster *t* masses was then compared against a null distribution for the cluster level *t* mass that was determined via the permutation. The permutation consisted of randomly permuting the true labels for the test data prior to computing accuracy. Since the labels are random, decoding accuracy remains at chance, and since we are assigning the same permuted label to each time point in a given epoch, we account for the temporal autocorrelation in the data during the construction of the empirical null distribution. For the permutations, clusters of *t* masses were found by summing the *t* scores of contiguous timepoints that were significant at $p < 0.05$. If there were no significant clusters of time points, then the cluster mass was set to zero for that permutation trial. If there were more than one cluster of significant time points, the mass of the largest cluster was taken. This whole permutation process was repeated 100 times across the 100 iterations of decoding accuracy. This resulted in the empirical null distribution of the cluster-level t mass, resulting in a resolution of $p = 0.001$. The *p* value corresponding to each cluster in the actual data set was computed by examining where each observed *t* mass fell within the null distribution. The *p* value for a given cluster was set based on the nearest percentiles of the null distribution. If the obtained cluster-level t mass was larger than the maximum of permuted cluster-level *t* mass, then the null hypothesis was rejected, and we concluded that the decoding was above chance for any observed cluster-level *t* mass that was in the top 95% of the null distribution (alpha = 0.05). For temporal

generalization this process was performed on a row-by-row basis, to assess the resulting time series that comprised the heat maps.

Computational Models

Optimal Bayesian Learner

We generated 1000 reward probability schedules with randomly generated magnitudes for each stimulus ranging from 1-100 points on a given trial. For each schedule we employed a Bayesian computational model to model an optimal participants' choice on a given trial (t) by considering choice history, outcome history, available choice options, and reward magnitudes; for extended details see (Behrens et al., 2007; Boorman et al., 2009, 2011). The associative contingencies were unobservable to the learner, so the model employed a Markovian approach to estimate the subjective belief that a given choice and a given counterfactual (C) would result in a given outcome and what would have happened in a counterfactual outcome (O), independently as follows:

$$p(S \rightarrow O) = q_s$$

Equation 1

q_s, denotes the probability of receiving a given outcome for either the choice or counterfactual, O, followed by a given choice S. On a trial-by-trial basis the model estimates the latest q_{st} value based on the observations from previous outcomes. The choice to outcome contingencies ranged from 0-1 and their likelihood is modeled as a beta distribution over possible values:

$$\beta(q_{st} | \, v)$$

Equation 2

The variance (v) of the beta distribution decides how likely (q_{st}) there is to be change on the next trial compared to the current trial. The variance also termed the volatility controlled the

extent to which observed outcomes updated choice-to-outcome contingencies on every trial, a large v suggests that q_{st} on the next trial would be more likely to change and the inverse is true as well. Volatility changes on a trial-to-trial basis as follows:

$$p(v_t|v_{t-1}, k) = N(v_{t-1}|k)$$

Equation 3

k denotes the rate of change of the volatility, v acting as the control parameter on the belief of how likely the rate of change changes. The distribution of v is modeled as Gaussian (N) in a Markovian manner. After each observation, the model updates choice-to-outcome contingencies only using the belief about the latest state of the contingency, q_{st-1}, and volatility, v_{t-1}. The prior, q_{st-1} captures the most likely state integrating the full history of past choices and outcomes without needing to store the history of the environment. The estimate of parameters therefore follows Bayes rule:

$$p(q_{s,t}, v_t, k|y_{1:t})$$
$$= p(y_t|q_{s,t}) \int [\int p(\, q_{s,t-1}, v_{t-1}|y_{1:t-1}) p(v_t|v_{t-1}) dv_{t-1}]\, p(q_{s,t}|q_{s,t-1}, vt) dq_{s,t-1}$$

Equation 4

This gives us a 3-dimensional grid of the joint probability distribution stored between trials. At the start of each run, the distribution is set to be uniform because participants are given no information about any of the parameters. On every trial the learner only needs to know the estimated choice-to-outcome and choice-to-counterfactual contingencies which both are marginalized over their respective v and k as follows:

$$p(q_{s,t}) = \int \int p(q_{s,t}, v_t, k) dv_t dk$$

Equation 5

The mean of the distribution can then be taken as follows:

$$\bar{q}_{s,t} = \int q_{s,t}\, p(q_{s,t}) dq_{s,t}$$

Equation 6

The model can then calculate the expected value of each stimulus on a given trial as follows:

$$EV_{s,t} = \bar{q}_{s,t}\, m_{O,t}\alpha$$

Equation 7

Where $m_{O,t}$ indicating the magnitude of reward on the current trial for a given outcome be it the choice or counterfactual and α denoting a free parameter that is the preference of either option.

$$p(choice = s) = \frac{e^{bEV_{s,t}}}{e^{bEV_{s,t}} + e^{bEV_{c,t}}}$$

Equation 8

s and c are the two stimuli on a given trial and b is the inverse temperature a free parameter that indicates the level of sensitivity of choices to expected values. Free parameters (α and b) were fit using Markov Chain Monte Carlos Methods. The parameter constraints were set to be: $0 < \alpha < 1$ and $0 < b < 7$. Initial values were set at $\alpha = 1$ (meaning no choice preference) and $b = 3.5$ (the mean of the inverse temperature bounds). The model was then fitted to maximize the likelihood of the optimal choice considering the trial-by-trial estimates of expected value as denoted in equation 8.

Of the 1,000 different schedules the Bayesian learner played, we chose a reward schedule where the learner learned a correlation between stimuli of $r < 0.2$ and learned a correlation of their independent magnitudes of < 0.2. We also had a constraint on sequential choices where the

learner could not repeat choosing a given stimuli > 3 times. Finally, learner's choices between stimuli were ensured to be balanced where out of a total 200 trials each stimulus was chosen ~100 times. This allowed us to use a reward schedule in our bandit task that would lead to balanced choices in theory. To confirm that participants' choices were in fact balanced and tracking onto the reward schedules, we fit three general linear models relating the learner's estimates, reward schedule magnitudes, and the outcomes of choice/counterfactual options against the participant's choices.

In the first model, we fit participants' choices to the expected value difference on a given trial using a general linear model with a logit link function as follows:

$$ln\left(\frac{p(chosen)}{p(unchosen)}\right) = \beta_0 + \beta_{\Delta EV}$$

General Linear Model 1

Using the beta weight and intercept, we then computed the logistic function for each participant and evaluated the predicted choice given the expected value difference on a trial-by-trial basis. If participants choices were balanced, then the indifference point between a participant choosing one option over the other, should be 0.5 when the expected value difference is 0.

In the second model we fit the participants' choices against the probability estimates ($pEst$) for either stimulus (1 or 2) and their respective magnitudes (m) as follows:

$$ln\left(\frac{p(chosen)}{p(unchosen)}\right) = \beta_0 + \beta_{pEst1} + \beta_{pEst2} + \beta_{m1} + \beta_{m2}$$

General Linear Model 2

With this model we assessed the beta weight of each probability estimate and magnitude to ensure that the participants were weighting the options independently of one another. If this was the case, then the weights for each option should go in inverse directions. Furthermore, if

participants were learning the probability distributions associated with either stimulus, then these betas should be larger than the betas associated with the magnitudes.

Our final model was used to assess behavioral learning effects on choices after participants received feedback for both stimuli (face (f) or house (h)) when they were either the chosen (C) or the counterfactual (CF) ranging from 1 – 3 trials back as follows:

$$\ln\left(\frac{\mathrm{p(chosen)}}{\mathrm{p(unchosen)}}\right) = \beta_0 + \beta_{Cf1} + \beta_{Cf2} + \beta_{Cf3} + \beta_{CFf1} + \beta_{CFf2} + \beta_{CFf3} + \beta_{Ch1} + \beta_{Ch2} + \beta_{Ch3} + \beta_{CFh1} + \beta_{CFh2} + \beta_{CFh3}$$

General Linear Model 3

If participants were integrating information about either stimulus independently, then the beta weights for a give stimulus should be showing an influence on the given trials choice in opposite directions with the most recent feedback for a given stimulus irrespective of chosen or unchosen should have the largest weight followed by a decay in influence for the subsequent outcomes.

Results

Behavior and Models

When looking at the overall choices participants made during the bandit task, there was no significant difference between choosing the face stimulus (M = 97.3, SD = 12.1) vs. the house stimulus (M = 102.2 SD = 12.1), indicating that choices were balanced across stimuli ($p = 0.364$, $t(20) = .929$). When we fit the participants choices against the expected value difference on a given trial, GLM 1, we saw that the indifference point was at y = 0.5, meaning that participants were balanced and optimal (Figure 1C). The model where we assessed the influence of the

probability estimates and magnitudes for either stimulus, GLM 2, showed that participants were weighting the probabilities of choosing either stimulus independently (Figure 1D). This is highlighted by the fact that the weights for either stimulus's estimates and magnitudes are on opposite sides of the axis. In our n-back model, GLM 3, we saw that for both the face and house, irrespective of whether the feedback was for the chosen option or the counterfactual, participants were integrating feedback information that influenced their current choice up to 3 trials back. Most weights for either model were significantly different from chance at a minimum of $p < 0.05$, and can be seen in Figure 1D.

SVM Decoding Results

Localizer Decoding

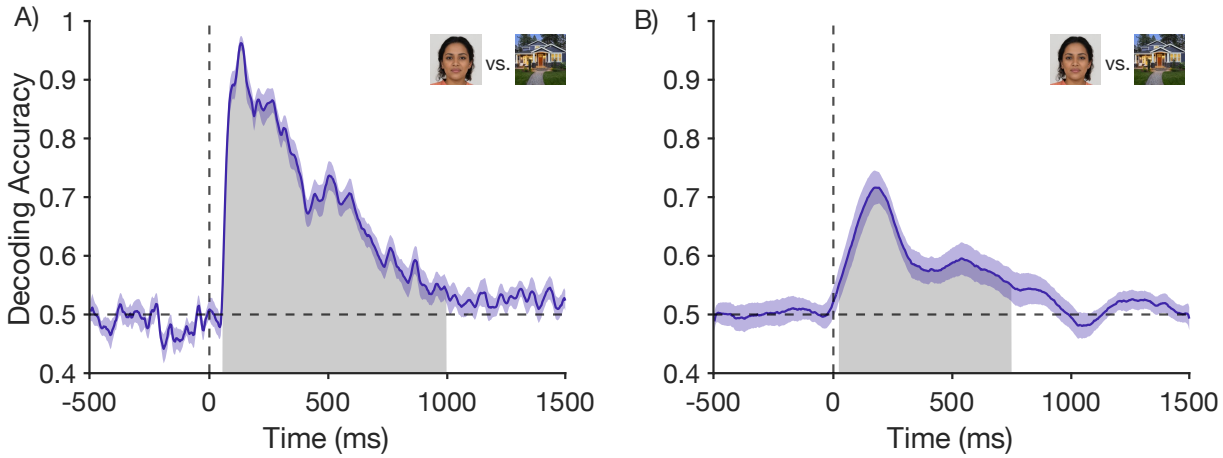


Figure 2. Decoding accuracy at each time point for the localizer task for the time-locked voltage (A) and the instantaneous alpha-band power (B). Shading indicates clusters of time points that were significantly different from chance after correction for multiple comparisons. For the time-locked voltage, the decoders were able to distinguish between the face and house stimuli for an extended period. For the alpha power, decoding accuracy was lower and was above chance for a shorter period. These results indicate that information about stimulus identity was represented in both signals.

Participants were at ceiling when performing the localizer task (89% accuracy), so we all the trials except those containing artifacts. Figure 1 shows the ability of the decoding approach to

distinguish between the face and house stimuli. As shown in Figure 1A, decoding accuracy for the time-locked voltage signal was significantly above chance for a long-time interval. Decoding accuracy peaked at 96% correct (at 132 ms) and remained significant until 1012 ms. For the alpha band power (Figure 2B), the decoding accuracy was similar but weaker. In this case, the peak decoding was lower at 72% correct (at 176 ms), and decoding remained significant until only 748 ms. This confirmed that we could indeed discriminated between the neural representations of the faces and houses with high fidelity using both the time-locked voltage and the instantaneous alpha power.

Bandit Task Decoding

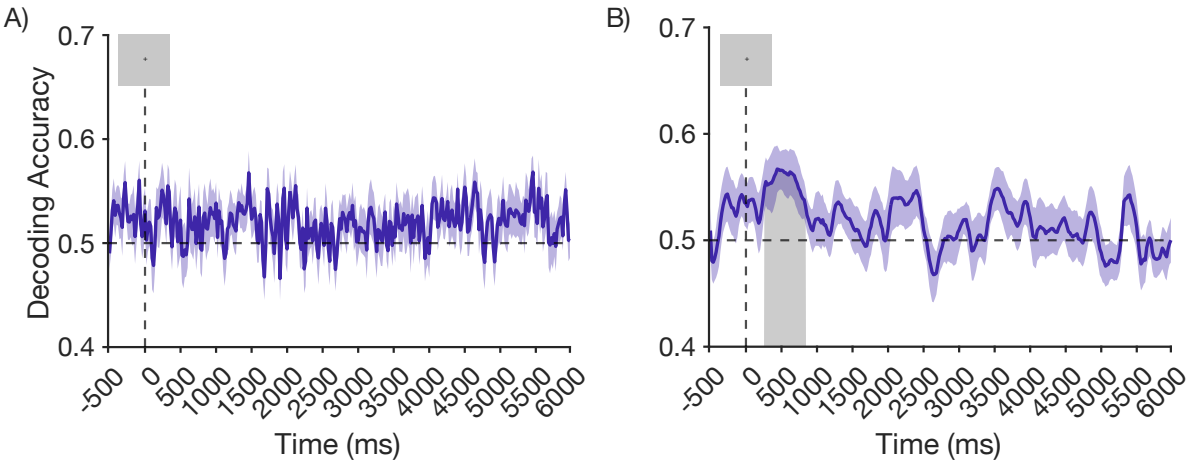


Figure 3. Decoding accuracy during the delay period following the choice for the time-locked voltage (A) and the instantaneous alpha power (B). Decoding accuracy for the time-locked voltage was near chance throughout the 6 second delay period, but alpha-band decoding accuracy was significantly greater than chance from 260 – 840 ms. This suggests that the choice representation is not actively maintained prior to feedback and thus must be reinstated.

We next attempted to decode whether the chosen stimulus was the face or the house during the 6000-ms delay period following the choice (Figure 3). For the time-locked voltage, decoding accuracy did not differ from chance at any point. Significant decoding was observed between 260-840 ms for the instantaneous alpha power, but this was not sustained until the end

of the delay. The lack of sustained decoding for either signal suggests that the stimuli or choices were not actively maintained across the delay and would therefore have to be reinstated at feedback.

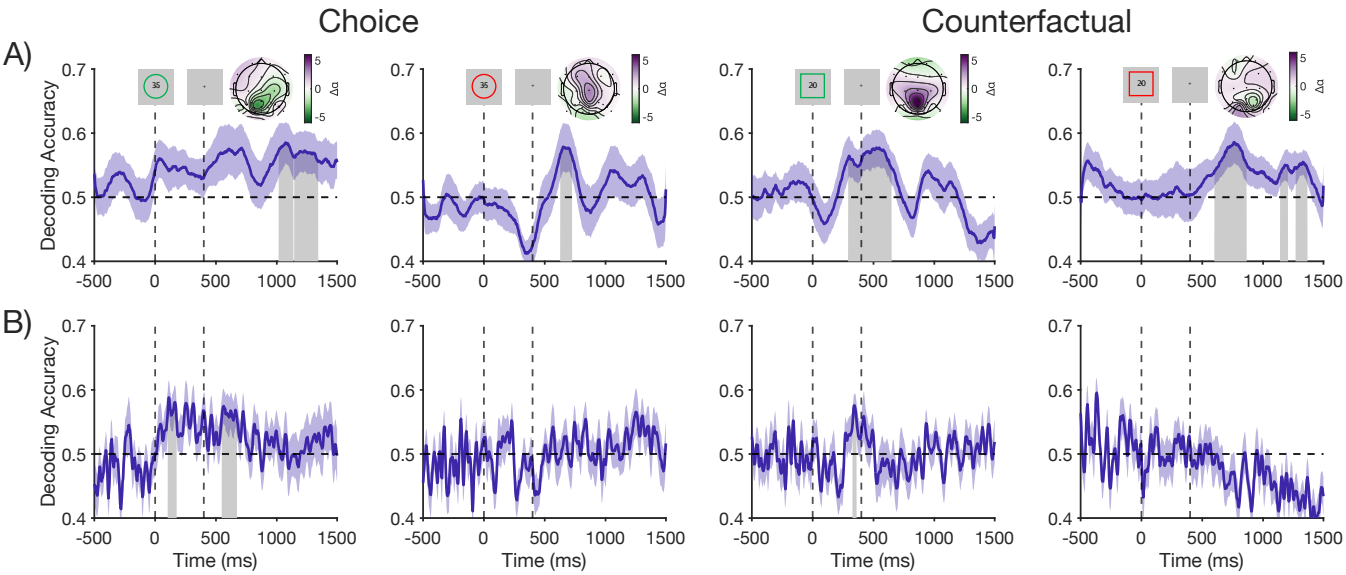


Figure 4. Decoding accuracy following the feedback stimulus for the chosen stimulus (Choice) and the unchosen stimulus (Counterfactual), separately when the feedback indicated that the stimulus was rewarded on that trial (or would have been rewarded if chosen in the case of the counterfactual). Alpha band decoding (A) showed a significant effect for all feedback screens, but during different latency bands. Win choice feedback showed significant decoding later in the epoch (1016-1336 ms) whereas choice losses showed significance earlier (628-724 ms). Counterfactual decoding showed a significant win effect (296-648 ms) that was earlier than the loss effect significance (600-864 ms). There were two additional brief periods of significance for the loss counterfactual at (1140-1204 ms) and (1264-1368 ms). Insets show univariate topographical alpha band power difference maps for feedback related to either stimulus within the main windows of significance for each decoding analysis. Time-locked voltage decoding (B) showed a few brief periods of significance for choice wins (100-180 ms; 532-680 ms) and counterfactual wins (328-360 ms) but was not significantly greater than chance for loss feedback. Clusters that are significantly different from chance are shown with gray shading.

Figure 4 shows decoding accuracy following the feedback for the chosen stimulus and for the unchosen (counterfactual) stimulus, separately when the feedback indicated a win for that stimulus or a loss for that stimulus. For the instantaneous alpha-band power (Figure 4A), stimulus identity could be decoded significantly about chance following the feedback for both wins and losses and for both the chosen stimulus (Choice) and the unchosen stimulus (Counterfactual). Choice decoding was significant late in the epoch from 1016-1336 ms for wins,

and significant earlier from 628-724 ms for losses. Counterfactual win decoding showed significance relatively early 296-648 ms, and counterfactual loss decoding showed significance from 600-864 ms, from 1140-1204 ms, and from 1264-1368 ms. Note that the periods of significance merely indicate that a decodable signal was present, and a lack of significance during a specific time period is not positive evidence for the lack of a signal during that period. Moreover, although the cluster mass statistical analysis maximizes power, it does not allow strong conclusions to be drawn about the timing of the effects (Groppe et al., 2011; Sassenhagen & Draschkow, 2019). Thus, it is important not to overinterpret the differences in the timing of the significant decoding effects across the four different types of feedback stimuli. Nonetheless, these results make it clear that information about the choice was present in the alpha-band signal following the feedback stimulus, even though we saw no evidence that this information was sustained throughout the delay period preceding the feedback. This provides evidence that the choice was reinstated following the feedback.

An inset image in the top right of each alpha band decoding plot shows the scalp map of the univariate alpha power difference between the face and house feedback signals within the main significance windows of the corresponding decoding analysis. These plots show that the difference was present primarily in the parietal and occipital electrode sites, where alpha-band oscillations are typically largest.

When we performed the same set of analyses using the time-locked voltage (Figure 4B), we saw a few brief periods of significance for choice wins from 100-180 ms and 532-680 ms but no significant decoding for choice losses. For counterfactual wins, we saw a brief period of significance from 328-360 ms, but counterfactual losses showed no significant decoding effects.

These results provide additional evidence that the choice was reinstated following the feedback, at least in the case of positive feedback.

Cross Task Temporal Generalization Decoding

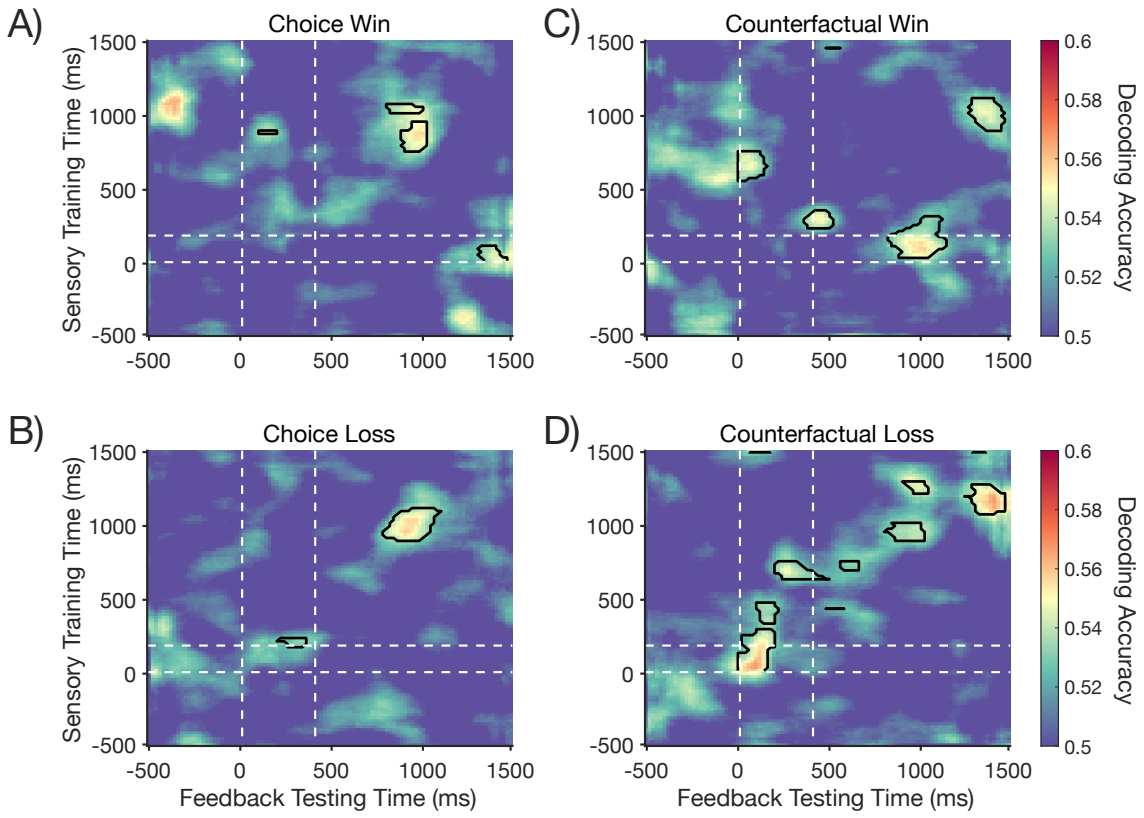


Figure 5. Cross-task, cross-temporal decoding accuracy during the feedback period of the bandit task. The X axis shows the latency being tested following the feedback, and the Y axis shows the latency of the localizer response used for training. Clusters of statistically significant decoding after correction for multiple comparisons are shown with black outline. Choice wins and losses (A, B) exhibited significant decoding when the decoder was trained ~1000 ms after stimulus onset in the localizer task and tested ~1000 ms after feedback onset. Counterfactual feedback wins and losses (C, D) also showed significant late effects, along with earlier clusters.

Figure 5A shows decoding accuracy for the cross-task, cross-temporal decoding analysis, where the decoder was trained on signals from the localizer task at a given time point and tested during the feedback epochs at every different time point. The goal of this analysis was to determine whether the representations assessed with the simple sensory discrimination task used

during the localizer are reinstated during the feedback period of the bandit task. Warmer colors indicate a higher decoding accuracy. Results are shown only for instantaneous alpha power; no significant effects were observed for the time-locked voltage.

For both choice wins and choice losses, decoding accuracy was significantly above chance when the decoder was trained ~1000 ms after stimulus onset in the localizer task and tested ~1000 ms after feedback onset. A similar late effect was observed for counterfactual wins, along with significant clusters when the decoder was trained at relatively early latencies in the localizer data (0-300 ms) and tested between approximately 500-1000 ms after the feedback stimulus in the bandit data. For counterfactual losses, significant decoding clusters were observed across the entire latency range corresponding to similar latencies in the localizer training data and the bandit test data. As noted above, it is important not to overinterpret the specific latency ranges of the significant clusters. The key finding was that significant cross-task decoding was present, indicating that the neural patterns that distinguished between the face and house stimuli during the localizer task were also present during the feedback period of the bandit task, consistent with reinstatement of the neural codes following the feedback stimuli, presumably to facilitate credit assignment.

Discussion

In this study we sought to provide evidence of the reinstatement of neural working memory representations at feedback to facilitate credit assignment. We had participants complete a localizer task that gave us sensory related signals to train a decoder and our novel 2-armed bandit task incentivized model-based credit assignment. We then used the time-locked voltages and the instantaneous alpha band power to assess the information content of the neural signals via decoding.

Participants' behavior showed that they could track the probability of reward as it drifted over time for each stimulus independently and that they were being primarily driven by the reward probabilities and not the randomly generated magnitudes (see the GLMs in Figure 1D). The Bayesian learner additionally ensured that the inferred probability estimates were independent from one another, and that participants' choices were balanced among the two stimuli (Figure 1C). This allowed us to be confident that the neural signals we were going to attempt to decode contained information related to participants' learning and decision-making.

In the localizer task, we could decode the neural representations of the face and house stimuli extremely well with the time-locked voltages (96% correct at the peak) and reasonably well with alpha band power (72% correct at the peak). The peak of the time-locked voltage decoding occurred earlier (132 ms) than the alpha band peak (176 ms), suggesting that the time-locked voltages are more sensitive to early sensory features and the alpha band activity primarily reflects higher-level attentional mechanisms (Bae & Luck, 2019; Bae & Luck, 2018; Noah et al., 2020).

We predicted that participants would not actively maintain the chosen stimulus in working memory during the relatively long (6000 ms) delay period following the choice, and this was supported by the decoding results during this period. Specifically, we found no significant decoding of the chosen stimulus from the time-locked voltages during this period, and only a brief period of significant alpha-band decoding during the initial portion of the delay (246-840 ms). This suggests that the choice was not maintained actively but was instead stored in an activity-silent state. Working memory representations are known to be able to be reactivated from these activity-silent states (Li et al., 2023; Printzlau et al., 2022; Rose et al., 2016).

We predicted that participants would show distinguishable information content about the neural representations associated with either face or house stimulus at both the choice and counterfactual outcome feedback. This prediction was supported by not only our decoding results, where we found significant effects for all feedback conditions but also by our computational models, specifically GLM 3 where we found evidence of participants integrating information about feedback up to 3 trials back to make their next choice for all feedback outcomes. This falls in line with previous choice and counterfactual work done with bandit tasks where there is evidence that it useful to integrate information about both (Boorman et al., 2011; Levens et al., 2014; Rushworth et al., 2011; Tobia et al., 2014)

A dominant hypothesis about alpha-band activity is *gating by inhibition*, where alpha is theorized to gate information via functional inhibition of task irrelevant areas (Jensen & Mazaheri, 2010). These selective attention mechanisms could have more larger fields of reach, across broad swaths of cortex, which could explain why we were only primarily able to decode the reinstatement of attentional mechanisms related to either stimulus at outcome or counterfactual feedback with the alpha power and not with the time-locked voltage. The timing of the significant periods of decoding varied across the different feedback types, but not in a systematic and interpretable manner. This is probably because the statistical power is lower for determining the specific times at which decoding was above chance than for determining whether decoding was simply above chance at some point in the epoch.

Finally, our cross-task, cross-temporal generalization decoding analysis indicated that the same alpha mechanisms used to perform a simple discrimination task (during the localizer task) are reinstated following feedback when credit assignment takes place. Previous research has found that multiple areas are activated in response to outcome and counterfactual feedback:

Outcome-related activity has been primarily seen in orbitofrontal cortex, dorsolateral prefrontal cortex, parietal cortex, anterior cingulate cortex, and hippocampus (Asaad et al., 2017; Noonan et al., 2017; Rushworth et al., 2011; Walton et al., 2010), whereas counterfactual-related activity has primarily been seen in the lateral frontopolar cortex, dorsomedial frontal cortex, posteromedial cortex, posterior parietal cortex, and hippocampus (Boorman et al., 2011; Parikh et al., 2018; Van Hoeck et al., 2015). Given the low spatial resolution of scalp EEG voltages, however, the present decoding results could reflect any combination of these cortical regions. What we can say—with greater temporal precision than is possible in previous research— is that information about the chosen stimulus is reinstated at feedback, presumably to support credit assignment.

Data and Code Availability

All data will be available upon publication.

Author Contributions

Carlos Daniel Carrasco: conceptualization, software, analyses, writing; Yanchang Zhang: conceptualization, software, writing; Steven J. Luck: conceptualization, writing, development of methods, funding acquisition. Erie D. Boorman: conceptualization, writing, development of methods, funding acquisition. Editing/Review was performed by all authors prior to submission of the final version of the manuscript.

Acknowledgements

We thank the members of the Learning and Decision-Making Lab for their valuable contributions to community and thoughtful insights, including but not limited to: Phil Witkowski, Seongmin (Alex) Park, Erica Varga, Mark Orloff, Lindsay Rondot, and Sarah Sweigart. We also thank Gi-Yeul Bae for his sharing of the temporal generalization code which we modified accordingly to fit this paradigm.

Funding

This study was supported by an early career award NSF grant presented to Erie. D. Boorman on which Steven J. Luck is a Co-PI.

References

Asaad, W. F., Lauro, P. M., Perge, J. A., & Eskandar, E. N. (2017). Prefrontal neurons encode a solution to the credit-assignment problem. *Journal of Neuroscience*, *37*(29), 6995–7007. https://doi.org/10.1523/JNEUROSCI.3311-16.2017

Bae, G. Y., & Luck, S. J. (2019). Decoding motion direction using the topography of sustained ERPs and alpha oscillations. *NeuroImage*, *184*, 242–255. https://doi.org/10.1016/j.neuroimage.2018.09.029

Bae, G. Y., & Luck, S. J. (2018). Dissociable Decoding of Spatial Attention and Working Memory from EEG Oscillations and Sustained Potentials. *The Journal of Neuroscience*, *38*(2), 2860–17. https://doi.org/10.1523/JNEUROSCI.2860-17.2017

Bae, G. Y., & Luck, S. J. (2019). *Appropriate Correction for Multiple Comparisons in Decoding of ERP Data: A Re-Analysis of Bae & Luck (2018)*. https://doi.org/10.1101/672741

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., & Rushworth, M. F. S. (2009). How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron*, *62*(5), 733–743. https://doi.org/10.1016/j.neuron.2009.05.014

Boorman, E. D., Behrens, T. E., & Rushworth, M. F. (2011). Counterfactual choice and learning in a Neural Network centered on human lateral frontopolar cortex. *PLoS Biology*, *9*(6). https://doi.org/10.1371/journal.pbio.1001093

Carrasco, C. D., Bahle, B., Simmons, A. M., & Luck, S. J. (2024). Using multivariate pattern analysis to increase effect sizes for event-related potential analyses. *Psychophysiology*. https://doi.org/10.1111/psyp.14570

Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, *21*(10), 576–586. https://doi.org/10.1038/s41583-020-0355-6

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

Collins, A. G. E., & Frank, M. J. (2018). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(10), 2502–2507. https://doi.org/10.1073/pnas.1720963115

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. In *Cognitive, Affective and Behavioral Neuroscience* (Vol. 14, Issue 2, pp. 473–492). Springer New York LLC. https://doi.org/10.3758/s13415-014-0277-8

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. https://doi.org/10.1016/j.conb.2008.08.003

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. In *Journal of Neuroscience Methods* (Vol. 134). http://www.sccn.ucsd.edu/eeglab/

Deserno, L., Moran, R., Michely, J., Lee, Y., Dayan, P., & Dolan, R. J. (2021). Dopamine enhances model-free credit assignment through boosting of retrospective model-based inference. *ELife*, *10*. https://doi.org/10.7554/eLife.67778

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., & Brea, J. (2018). Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. In *Frontiers in Neural Circuits* (Vol. 12). Frontiers Media S.A. https://doi.org/10.3389/fncir.2018.00053

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. In *Psychophysiology* (Vol. 48, Issue 12, pp. 1711–1725). Blackwell Publishing Inc. https://doi.org/10.1111/j.1469-8986.2011.01273.x

He, K., Huertas, M., Hong, S. Z., Tie, X. X., Hell, J. W., Shouval, H., & Kirkwood, A. (2015). Distinct Eligibility Traces for LTP and LTD in Cortical Synapses. *Neuron*, *88*(3), 528–538. https://doi.org/10.1016/j.neuron.2015.09.037

Jensen, O., & Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: Gating by inhibition. *Frontiers in Human Neuroscience*, *4*. https://doi.org/10.3389/fnhum.2010.00186

Levens, S. M., Larsen, J. T., Bruss, J., Tranel, D., Bechara, A., & Mellers, B. A. (2014). What might have been? The role of the ventromedial prefrontal cortex and lateral orbitofrontal cortex in counterfactual emotions and choice. *Neuropsychologia*, *54*(1), 77–86. https://doi.org/10.1016/j.neuropsychologia.2013.10.026

Li, D., Hu, Y., Qi, M., Zhao, C., Jensen, O., Huang, J., & Song, Y. (2023). Prioritizing flexible working memory representations through retrospective attentional strengthening. *NeuroImage*, *269*. https://doi.org/10.1016/j.neuroimage.2023.119902

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*(1 APR), 1–14. https://doi.org/10.3389/fnhum.2014.00213

MathWorks. (2023). *MATLAB version: 23.2.0 (R2023b)*. The MathWorks Inc. https://www.mathworks.com

Moran, R., Dayan, P., Dolan, R. J., Moran, R., Dayan, P., & Dolan, R. J. (2021). Article Efficiency and prioritization of inference-based credit assignment Efficiency and prioritization of inference-based credit assignment. *Current Biology*, 1–10. https://doi.org/10.1016/j.cub.2021.03.091

Moran, R., Keramati, M., Dayan, P., & Dolan, R. J. (2019). Retrospective model-based inference guides model-free credit assignment. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-08662-8

Nadra, J. G., Bengson, J. J., Morales, A. B., & Mangun, G. R. (2023). Attention without Constraint: Alpha Lateralization in Uncued Willed Attention. *ENeuro*, *10*(6). https://doi.org/10.1523/ENEURO.0258-22.2023

Noah, S., Powell, T., Khodayari, N., Olivan, D., Ding, M., & Mangun, G. R. (2020). Neural Mechanisms of Attentional Control for Objects: Decoding EEG Alpha When Anticipating Faces, Scenes, and Tools. *Journal of Neuroscience*, *40*(25), 4913–4924. https://doi.org/10.1523/JNEUROSCI.2685-19.2020

Noonan, M. P., Chau, B. K. H., Rushworth, M. F. S., & Fellows, L. K. (2017a). Contrasting effects of medial and lateral orbitofrontal cortex lesions on credit assignment and decision-making in humans. *Journal of Neuroscience*, *37*(29), 7023–7035. https://doi.org/10.1523/JNEUROSCI.0692-17.2017

Noonan, M. P., Chau, B. K. H., Rushworth, M. F. S., & Fellows, L. K. (2017b). Contrasting effects of medial and lateral orbitofrontal cortex lesions on credit assignment and decision-making in humans. *Journal of Neuroscience*, *37*(29), 7023–7035. https://doi.org/10.1523/JNEUROSCI.0692-17.2017

NVIDIA. (2019). *Face Generator – Generate Faces Online Using AI — generated.photos*.

O'toole, A. J., Jiang, F., Abdi, H., & Haxby, J. V. (2005). *Partially Distributed Representations of Objects and Faces in Ventral Temporal Cortex*. http://mitprc.silverchair.com/jocn/article-pdf/17/4/580/1757195/0898929053467550.pdf

Pan, W. X., Schmidt, R., Wickens, J. R., & Hyland, B. I. (2005). Dopamine cells respond to predicted events during classical conditioning: Evidence for eligibility traces in the reward-learning network. *Journal of Neuroscience*, *25*(26), 6235–6242. https://doi.org/10.1523/JNEUROSCI.1478-05.2005

Parikh, N., Ruzic, L., Stewart, G. W., Spreng, R. N., & De Brigard, F. (2018). What if? Neural activity underlying semantic and episodic counterfactual thinking. *NeuroImage*, *178*, 332–345. https://doi.org/10.1016/j.neuroimage.2018.05.053

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Printzlau, F. A. B., Myers, N. E., Manohar, S. G., & Stokes, M. G. (2022). Neural reinstatement tracks spread of attention between object features in working memory. *Journal of Cognitive Neuroscience*, *34*(9), 1681–1701. https://doi.org/10.1162/jocn_a_01879

Richards, B. A., & Lillicrap, T. P. (2019). Dendritic solutions to the credit assignment problem. In *Current Opinion in Neurobiology* (Vol. 54, pp. 28–36). Elsevier Ltd. https://doi.org/10.1016/j.conb.2018.08.003

Rose, N. S., Larocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136–1140.

Rushworth, M. F. S., Noonan, M. A. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal Cortex and Reward-Guided Learning and Decision-Making. *Neuron*, *70*(6), 1054–1069. https://doi.org/10.1016/j.neuron.2011.05.014

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*, *56*(6). https://doi.org/10.1111/psyp.13335

Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research*, *76*, 31–42. https://doi.org/10.1016/j.visres.2012.10.012

Tobia, M. J., Guo, R., Schwarze, U., Boehmer, W., Gläscher, J., Finckh, B., Marschner, A., Büchel, C., Obermayer, K., & Sommer, T. (2014). Neural systems for choice and valuation with counterfactual learning signals. *NeuroImage*, *89*, 57–69. https://doi.org/10.1016/j.neuroimage.2013.11.051

Van Hoeck, N., Watson, P. D., & Barbey, A. K. (2015). Cognitive neuroscience of human counterfactual reasoning. *Frontiers in Human Neuroscience*, *9*(JULY), 1–18. https://doi.org/10.3389/fnhum.2015.00420

Walton, M. E., Behrens, T. E. J., Buckley, M. J., Rudebeck, P. H., & Rushworth, M. F. S. (2010). Separable Learning Systems in the Macaque Brain and the Role of Orbitofrontal Cortex in Contingent Learning. *Neuron*, *65*(6), 927–939. https://doi.org/10.1016/j.neuron.2010.02.027

Witkowski, P. P., Park, S. A., & Boorman, E. D. (2022). Neural mechanisms of credit assignment for inferred relationships in a structured world. *Neuron*, *110*(16), 2680-2690.e9. https://doi.org/10.1016/j.neuron.2022.05.021

Yoo, A. H., & Collins, A. G. E. (2022). How Working Memory and Reinforcement Learning Are Intertwined: A Cognitive, Neural, and Computational Perspective. *Journal of Cognitive Neuroscience*, *34*(4), 551–568. https://doi.org/10.1162/jocn_a_01808

Chapter 5: General Discussion

Science is incremental, and the body of work presented here is a steppingstone for others to continue unraveling and discovering various insights into the brain and cognitive processes. The first empirical chapter critically assessed whether multivariate pattern classification would yield larger signal to noise ratios than univariate approaches for several classic ERP components. In general, we found that multivariate techniques can yield as good and often larger effect sizes compared to traditional grand averaging. However, the consequence of being able to leverage the topographical ERP distribution of the scalp, rather than focusing on a specific site or cluster of sites related to a particular component, is a loss of specificity in terms of states of processing. Specifically, information about polarity is lost, which can often help isolate different neural processes. SVM decoding in general provides a metric of the overall information content of the neural signals irrespective of their polarity. Something to note is that when all the same analyses were repeated using clusters of sites relevant to a specific component rather than all sites, the same general pattern of decoding results were found across the time series – they yielded only slightly smaller effect sizes. Therefore, for these types of paradigms, it is likely that large univariate signals end up dominating the primary ongoing processes that can be detected across the scalp.

Decoding is often used for two primary cases in relation to the EEG: decoding for prediction or decoding for scientific understanding (Hebart & Baker, 2018). We primarily focused on the latter in this body of work, but decoding for prediction is a blooming complementary field that is approaching ceiling effects via optimization of largely neural network-based classification algorithms. In the early 2020s, transformer models have become popular in the creation of chatbots that provide human-like responses to various queries ranging from storytelling to programming. These models have been co-opted by researchers for

classification tasks; for example, one of the most benchmarked EEG datasets, a 4-class decoding problem, is approaching ceilings effects where in some cases researchers are achieving greater than 90% accuracy on classification of single trials (Arpaia et al., 2022; Hou et al., 2020; Liu et al., 2023; Song et al., 2023). Elsewhere, some are attempting to create EEG/MEG foundation models which provide large-scale pretrained networks that in theory help yield high accuracy for brain computer interfaces that are capable of classification out of the box (Chen et al., 2024; Csaky et al., 2024; Cui et al., 2023; Panchavati & Speier, 2024). Meanwhile, decoding for scientific understanding has seemed to converge on the use of predominantly linear models, for more direct interpretation of data, and the use of neural network models to look at the theoretical representation of content in various processing streams such as the ventral visual stream with various behavioral and neural measures (Bonnen et al., 2021; Grootswagers et al., 2017; Haufe et al., 2014; Kiat et al., 2022; Park et al., 2023). Notably, these advances in decoding are not limited to just cognitive neuroscience but are leveraged across a wide variety of fields, such as in biophysics where protein folding is now discovered via large scale computationally expensive deep learning networks (Casares et al., 2022; Lee et al., 2019; Noé et al., 2020; Xu, 2019).

Theories of working memory, ranging from low-level cell-based circuit models to high-level cognitive models, continue to be extensively studied and debated (Baddeley, 2010; Barak & Tsodyks, 2014; Chai et al., 2018; Cotton & Ricker, 2022; Kamiński & Rutishauser, 2020; Luck & Vogel, 2013; Miller et al., 2018). The second chapter of empirical work provided some of the first direct evidence that memory representations for rare events in visuospatial change detection tasks are enhanced. Behavioral analyses indicated that the representations of rare stimuli were more precise and that their degree of precision was correlated with the amplitude of the P3b component. Moreover, the identity rare stimuli could be decoded more accurately than

the identity of frequent stimuli. Thus, this study provided multiple sources of evidence that memory of rare events is enhanced. It will take follow-ups to replicate, validate, and extend our findings, but the task design should serve as a useful frame of reference for others to be able to further probe these processes. Circular stimulus spaces like the one used in this study are advantageous because they are continuous but do not suffer from edge artifacts (i.e., guessing strategies based on the possible range of stimulus values).

Developing algorithms that can extract more information from the electrophysiological recordings will be useful in creating more sophisticated working memory experiments relating information content to precise behavior on the level of single trials. One could imagine leveraging one of the prediction-based deep learning algorithms used in brain-computer interfaces, but further work is needed to figure out how to do this properly and with clear scientific interpretation. Initial implementations that have showed promise include deep learning full epoch decoding models that have been applied to EEG/MEG data (Csaky et al., 2023; Défossez et al., 2023; Hultén et al., 2021; Schirrmeister et al., 2017). We may be at a paradigm shift where the high dimensional nature of neural data is becoming less and less of a barrier. Technology continues to advance, such as the creation of > 3 tesla MRI scanners, dry electrodes for EEG, MEG based on optically pumped magnetometers, neuropixel recordings, optogenetics, fNIRS, functional ultrasound imaging, etc. (Boyden et al., 2005; Deffieux et al., 2021; Ferrari & Quaresima, 2012; Gutteling et al., 2023; Jun et al., 2017; Laine et al., 2019; Rabut et al., 2020; Van Der Kolk et al., 2013; Xu et al., 2017). These advancements will allow further refinement and specificity of our understanding.

In the last empirical chapter, we focused on credit assignment processes and their relationship with the working memory system. While in the chapter we operationalized the

working memory system as being a suitable candidate system for eliciting model-based credit assignment, the reality is that the way that credit assignment is dichotomized has been recently contested (Collins & Cockburn, 2020; Feher da Silva et al., 2023; Yoo & Collins, 2022). The muddled effects researchers often find of both model-free and model-based systems being engaged in learning tasks, often in hybridized ways, is potentially not suitable for dichotomization; rather, we should focus more on the systems that facilitate them, such as the reinforcement system and the working memory system, as it is the case that there is no system in the brain that independently facilitates these two formalizations of credit assignment on its own. In contrast to this, the combined ability of the reinforcement learning system and the working memory system to promote contingency learning representations is now beginning to be critically examined, and mechanistic models of how these processes work together are being explored (Chatham et al., 2014; Frank et al., 2001; Hazy et al., 2007; Traylor et al., 2024; Westbrook et al., 2024). Nonetheless, we saw evidence for credit assignment, whether it be model based or model free, being facilitated by both the reinstatement of the chosen/counterfactual stimuli following feedback, which highlights the pivotal nature of the working memory system in allowing for these retrospective processes to happen. Deep learning architectures such as those in recurrent neural networks, variable auto encoder/decoders, and transformers all have some sort of module that allows them to retain information over periods of time or directs ‘attention’ towards specific relevant inputs for performing some process like working memory, but these agents still do not have a system with this architecture. I suspect that future work in deep learning research, particularly in robotics, will eventually converge on working memory processes that will allow for composite interactions and allow the orchestration and integration of various processes to allow for increasingly human-like outputs.

Challenges involving the philosophical aspects of human cognition and the criteria for determining whether artificial intelligence (AI) has reached a level comparable to human capabilities are still beyond our grasp. Still, if human intelligence is not due to domain-specific evolutionary changes, as new theoretical work suggests, but rather differences in learning and memory capacities, then we might not be that far off (Cantlon & Piantadosi, 2024). As agents receive more and more pressure from their environment to perform increasingly complex tasks, humans themselves will evolve the capacities of these agents via the creation of cognitive inspired modules. This is already the case to some extent, as the architectures of the best deep learning image classification algorithms are all largely inspired by the ventral visual stream (He et al., 2016; Huang et al., 2017; Krizhevsky et al., 2012; Simonyan & Zisserman, 2014; Szegedy et al., 2015). Additionally, the best generative AI models to date were all developed based on our understanding of linguistics (Boleda, 2020; Pater, 2019; Pavlick, 2021, 2022). The concept of neural networks begets itself.

Given the state of the world where a majority of developed nations have entered or are entering a phase of hyper capitalism with ever expanding growth not being restricted despite our finite resources (Fanning et al., 2022; Jackson, 2019; Jackson & Victor, 2021), the monetary drive behind continual growth should accelerate the progress we see over the remainder of the 2020s. The nature of this society permeates throughout all major entities (corporations, government bodies, educational institutions, etc.) so there will be an ever-increasing expansion of more PhDs and more papers being put out, and thus more information growth. Recently, NVIDIA broke Moore's law with their new series of graphical processing units, exceeding the predicted doubling of computational power every two years (Alpha, 2022; Moore, 1998; NVIDIA, 2024; NVIDIA News, 2024). In turn there has been an increasing number of audio-

visual deepfakes, misuse of generative AI for the creation of adult content, the implementation of humanlike chatbots in almost every corner of the web related to customer support, expanses in automation, and the loss of a variety of jobs ranging from low to high skill level professions is eminent. Policy and legislative bodies have been trying to gain a grip on how to contain the immense expansion of AI, but Pandora's box has been opened; perhaps all we can do now is learn to deal with the new miseries of our world.

Still, we are not quite yet at that junction of impending doom, and I am hopeful that those in the pursuit of knowledge will continue to act in good faith. To have faith in humanity is to believe that the intrinsic drives that motivate humans, despite societal pressures, are predominantly good. Prosocial behaviors among humans have allowed for the creation of culture, society, large nation states and all advancement henceforth. Neuroscience is not pursued in a vacuum. There is a need in the study of neuroscience to integrate various fields, and often the most pivotal work is done by teams of scientists working together across various institutions and/or subfields. Ultimately the 21st Century will be a time of great expanse in all domains. If we can make it through the constant turmoil of violence, injustice, inequality, environmental destruction, and hate, then I am sure we can approach a world that stands on the basis of love, compassion, egalitarianism, and peace. Until then, we must continue to do our best to act in good faith in not only our science, but throughout all aspects of life.

References

Alpha, S. (2022). *NVIDIA: Growth Beyond Moore's Law*.

Arpaia, P., Esposito, A., Natalizio, A., & Parvis, M. (2022). How to successfully classify EEG in motor imagery BCI: A metrological analysis of the state of the art. In *Journal of Neural Engineering* (Vol. 19, Issue 3). Institute of Physics. https://doi.org/10.1088/1741-2552/ac74e0

Baddeley, A. (2010). Working memory. *Current Biology*, *20*(4), 136–140. https://doi.org/10.1016/j.cub.2009.12.014

Barak, O., & Tsodyks, M. (2014). Working models of working memory. In *Current Opinion in Neurobiology* (Vol. 25, pp. 20–24). https://doi.org/10.1016/j.conb.2013.10.008

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, *6*, 213–234.

Bonnen, T., Yamins, D. L. K., & Wagner, A. D. (2021). When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Neuron*, *109*(17), 2755-2766.e6. https://doi.org/10.1016/j.neuron.2021.06.018

Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G., & Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience*, *8*(9), 1263–1268. https://doi.org/10.1038/nn1525

Cantlon, J. F., & Piantadosi, S. T. (2024). Uniquely human intelligence arose from expanded information capacity. *Nature Reviews Psychology*, *3*(4), 275–293. https://doi.org/10.1038/s44159-024-00283-3

Casares, P. A. M., Campos, R., & Martin-Delgado, M. A. (2022). QFold: Quantum walks and deep learning to solve protein folding. *Quantum Science and Technology*, *7*(2). https://doi.org/10.1088/2058-9565/ac4f2f

Chai, W. J., Abd Hamid, A. I., & Abdullah, J. M. (2018). Working memory from the psychological and neurosciences perspectives: A review. In *Frontiers in Psychology* (Vol. 9, Issue MAR). Frontiers Media S.A. https://doi.org/10.3389/fpsyg.2018.00401

Chatham, C. H., Frank, M. J., & Badre, D. (2014). Corticostriatal output gating during selection from working memory. *Neuron*, *81*(4), 930–942. https://doi.org/10.1016/j.neuron.2014.01.002

Chen, Y., Ren, K., Song, K., Wang, Y., Wang, Y., Li, D., & Qiu, L. (2024). *EEGFormer: Towards Transferable and Interpretable Large-Scale EEG Foundation Model*. http://arxiv.org/abs/2401.10278

Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, *21*(10), 576–586. https://doi.org/10.1038/s41583-020-0355-6

Cotton, K., & Ricker, T. J. (2022). Examining the relationship between working memory consolidation and long-term consolidation. In *Psychonomic Bulletin and Review* (Vol. 29, Issue 5, pp. 1625–1648). Springer. https://doi.org/10.3758/s13423-022-02084-2

Csaky, R., van Es, M. W. J., Jones, O. P., & Woolrich, M. (2023). Interpretable many-class decoding for MEG. *NeuroImage*, *282*. https://doi.org/10.1016/j.neuroimage.2023.120396

Csaky, R., van Es, M. W. J., Jones, O. P., & Woolrich, M. (2024). Foundational GPT Model for MEG. *ArXiv Preprint ArXiv:2404.09256*.

Cui, W., Jeong, W., Thölke, P., Medani, T., Jerbi, K., Joshi, A. A., & Leahy, R. M. (2023). *Neuro-GPT: Towards A Foundation Model for EEG*. http://arxiv.org/abs/2311.03764

Deffieux, T., Demené, C., & Tanter, M. (2021). Functional Ultrasound Imaging: A New Imaging Modality for Neuroscience. In *Neuroscience* (Vol. 474, pp. 110–121). Elsevier Ltd. https://doi.org/10.1016/j.neuroscience.2021.03.005

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., & King, J. R. (2023). Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, *5*(10), 1097–1107. https://doi.org/10.1038/s42256-023-00714-5

Fanning, A. L., O'Neill, D. W., Hickel, J., & Roux, N. (2022). The social shortfall and ecological overshoot of nations. *Nature Sustainability*, *5*(1), 26–36. https://doi.org/10.1038/s41893-021-00799-z

Feher da Silva, C., Lombardi, G., Edelson, M., & Hare, T. A. (2023). Rethinking model-based and model-free influences on mental effort and striatal prediction errors. *Nature Human Behaviour*, *7*(6), 956–969. https://doi.org/10.1038/s41562-023-01573-1

Ferrari, M., & Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. In *NeuroImage* (Vol. 63, Issue 2, pp. 921–935). https://doi.org/10.1016/j.neuroimage.2012.03.049

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cognitive, Affective, & Behavioral Neuroscience*, *1*(2), 137–160.

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Gutteling, T. P., Bonnefond, M., Clausner, T., Daligault, S., Romain, R., Mitryukovskiy, S., Fourcault, W., Josselin, V., Le Prado, M., Palacios-Laloy, A., Labyt, E., Jung, J., & Schwartz, D. (2023). A New Generation of OPM for High Dynamic and Large Bandwidth MEG: The 4He OPMs—First Applications in Healthy Volunteers. *Sensors*, *23*(5). https://doi.org/10.3390/s23052801

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: Computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions*

of the Royal Society B: Biological Sciences, *362*(1485), 1601–1613. https://doi.org/10.1098/rstb.2007.2055

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. In *NeuroImage* (Vol. 180, pp. 4–18). Academic Press Inc. https://doi.org/10.1016/j.neuroimage.2017.08.005

Hou, Y., Zhou, L., Jia, S., & Lun, X. (2020). A novel approach of decoding EEG four-class motor imagery tasks via scout ESI and CNN. *Journal of Neural Engineering*, *17*(1). https://doi.org/10.1088/1741-2552/ab4af6

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708.

Hultén, A., van Vliet, M., Kivisaari, S., Lammi, L., Lindh-Knuutila, T., Faisal, A., & Salmelin, R. (2021). The neural representation of abstract words may arise through grounding word meaning in language itself. *Human Brain Mapping*, *42*(15), 4973–4984. https://doi.org/10.1002/hbm.25593

Jackson, T. (2019). The Post-growth Challenge: Secular Stagnation, Inequality and the Limits to Growth. *Ecological Economics*, *156*, 236–246. https://doi.org/10.1016/j.ecolecon.2018.10.010

Jackson, T., & Victor, P. A. (2021). Confronting inequality in the "new normal": Hyper-capitalism, proto-socialism, and post-pandemic recovery. *Sustainable Development*, *29*(3), 504–516. https://doi.org/10.1002/sd.2196

Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., Lee, A. K., Anastassiou, C. A., Andrei, A., Aydin, Ç., Barbic, M., Blanche, T. J., Bonin, V., Couto, J., Dutta, B., Gratiy, S. L., Gutnisky, D. A., Häusser, M., Karsh, B., … Harris, T. D. (2017). Fully integrated silicon probes for high-density recording of neural activity. *Nature*, *551*(7679), 232–236. https://doi.org/10.1038/nature24636

Kamiński, J., & Rutishauser, U. (2020). Between persistently active and activity-silent frameworks: novel vistas on the cellular basis of working memory. In *Annals of the New York Academy of Sciences* (Vol. 1464, Issue 1, pp. 64–75). Blackwell Publishing Inc. https://doi.org/10.1111/nyas.14213

Kiat, J. E., Luck, S. J., Beckner, A. G., Hayes, T. R., Pomaranski, K. I., Henderson, J. M., & Oakes, L. M. (2022). Linking patterns of infant eye movements to a neural network model of the ventral stream using representational similarity analysis. *Developmental Science*, *25*(1). https://doi.org/10.1111/desc.13155

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

Laine, P., Nenonen, J., Chappell, S., & Knuutila, J. (2019). Zero Helium Boiloff MEG Technology. In *Magnetoencephalography* (pp. 1–11). Springer International Publishing. https://doi.org/10.1007/978-3-319-62657-4_79-1

Lee, H., Turilli, M., Jha, S., Bhowmik, D., Ma, H., & Ramanathan, A. (2019). DeepDriveMD: Deep-learning driven adaptive molecular simulations for protein folding. *Proceedings of DLS 2019: Deep Learning on Supercomputers - Held in Conjunction with SC 2019: The International Conference for High Performance Computing, Networking, Storage and Analysis*, 12–19. https://doi.org/10.1109/DLS49591.2019.00007

Liu, K., Yang, M., Yu, Z., Wang, G., & Wu, W. (2023). FBMSNet: A Filter-Bank Multi-Scale Convolutional Neural Network for EEG-Based Motor Imagery Decoding. *IEEE Transactions on Biomedical Engineering*, *70*(2), 436–445. https://doi.org/10.1109/TBME.2022.3193277

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400.

Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. In *Neuron* (Vol. 100, Issue 2, pp. 463–475). Cell Press. https://doi.org/10.1016/j.neuron.2018.09.023

Moore, G. E. (1998). Cramming more components onto integrated circuits. *Proceedings of the IEEE*, *86*(1), 82–85.

Noé, F., De Fabritiis, G., & Clementi, C. (2020). Machine learning for protein folding and dynamics. In *Current Opinion in Structural Biology* (Vol. 60, pp. 77–84). Elsevier Ltd. https://doi.org/10.1016/j.sbi.2019.12.005

NVIDIA. (2024). *NVIDIA Blackwell Architecture*.

NVIDIA News. (2024). *NVIDIA Blackwell Platform Arrives to Power a New Era of Computing*.

Panchavati, S., & Speier, W. (2024). Mentality: A Mamba-Based Approach Towards Foundation Models for EEG. *ICLR 2024 Workshop on Learning from Time Series For Health*.

Park, S. A., Zolfaghar, M., Russin, J., Miller, D. S., O'Reilly, R. C., & Boorman, E. D. (2023). The representational geometry of cognitive maps under dynamic cognitive control. *BioRxiv*, 2022–2023.

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, *95*(1), e41–e74. https://doi.org/10.1353/lan.2019.0009

Pavlick, E. (2021). *Annual Review of Linguistics Semantic Structure in Deep Learning Keywords*. https://doi.org/10.1146/annurev-linguistics-031120

Pavlick, E. (2022). Semantic structure in deep learning. *Annual Review of Linguistics*, *8*, 447–471.

Rabut, C., Yoo, S., Hurt, R. C., Jin, Z., Li, H., Guo, H., Ling, B., & Shapiro, M. G. (2020). Ultrasound Technologies for Imaging and Modulating Neural Activity. In *Neuron* (Vol. 108, Issue 1, pp. 93–110). Cell Press. https://doi.org/10.1016/j.neuron.2020.09.003

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, *38*(11), 5391–5420. https://doi.org/10.1002/hbm.23730

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.

Song, Y., Zheng, Q., Liu, B., & Gao, X. (2023). EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 710–719. https://doi.org/10.1109/TNSRE.2022.3230250

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Traylor, A., Merullo, J., Frank, M. J., & Pavlick, E. (2024). *Transformer Mechanisms Mimic Frontostriatal Gating Operations When Trained on Human Working Memory Tasks*. http://arxiv.org/abs/2402.08211

Van Der Kolk, A. G., Hendrikse, J., Zwanenburg, J. J. M., Visser, F., & Luijten, P. R. (2013). Clinical applications of 7 T MRI in the brain. In *European Journal of Radiology* (Vol. 82, Issue 5, pp. 708–718). Elsevier Ireland Ltd. https://doi.org/10.1016/j.ejrad.2011.07.007

Westbrook, A., van den Bosch, R., Hofmans, L., Papadopetraki, D., Maatta, J. I., Collins, A. G. E., Frank, M. J., & Cools, R. (2024). Striatal Dopamine Can Enhance Learning, Both Fast and Slow, and Also Make it Cheaper. *BioRxiv*, 2022–2024.

Xu, J. (2019). Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(34), 16856–16865. https://doi.org/10.1073/pnas.1821309116

Xu, J., Mitra, S., Van Hoof, C., Yazicioglu, R. F., & Makinwa, K. A. A. (2017). Active Electrodes for Wearable EEG Acquisition: Review and Electronics Design Methodology. In *IEEE Reviews in Biomedical Engineering* (Vol. 10, pp. 187–198). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/RBME.2017.2656388

Yoo, A. H., & Collins, A. G. E. (2022). How Working Memory and Reinforcement Learning Are Intertwined: A Cognitive, Neural, and Computational Perspective. *Journal of Cognitive Neuroscience*, *34*(4), 551–568. https://doi.org/10.1162/jocn_a_01808