Predictive approaches for investigating brain activity underlying successful learning

by

Sucheta Chakravarty

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Psychology
University of Alberta

Abstract

Successful learning is of vital importance to human cognition. Accordingly, researchers have been interested to understand brain-activity signals that support it. However, traditional analysis of brain activity is based on planned comparisons and descriptive methods, which can both overestimate brain activity by overfitting it, and also underestimate the behavioural relevance of brain-activity measures by ignoring the subtle multivariate patterns. Complementary to the traditional analysis, here I used predictive approaches that can offer a stronger framework for finding behaviourally-relevant brain activity by testing predictions for unseen/future observations. Two learning situations were considered; one, where participants studied lists of words followed by old/new recognition tests for target (studied) and lure (new) items, the other involved trial-and-error learning of the stimulus-response rules for a large set of words, driven by reward feedback. For both learning paradigms, I asked if brain activity present when participants studied the material explained subsequent variability in the learning outcomes. First, I tested if features of brain-activity signals present during the study phase, as identified by previous planned-comparisons based investigations, could withstand tests of predictions for learning outcomes at the level of individual trials. For both tasks, this produced a small but significant success across a large number of participants. Next, I asked if data-driven multivariate pattern analysis of the study-phase activity produced better predictions for the learning outcomes. The multivariate pattern analysis achieved a small significant success for the item-recognition task, but it was under-powered for the trial-and-error learning task and produced non-significant results. Taken together, for both tasks, the contribution of the study-phase activity to later variability in learning outcomes was overall small, indicating that other important predictors

may be missing. To test this suggestion, I further investigated brain activity during the test phase of the item recognition task. Following a similar approach as study, first I used features of previously-identified brain-activity signals, to predict the memory outcomes, which achieved modest success. Then, I conducted multivariate pattern analysis of test-phase activity, which was still modest but predicted significantly better than the individual signals at study or test phases, as well as the multivariate activity at study. Further, combining brain-activity features from both study and test phases led to similar size of predictions as that for the test-phase only. Thus, test-phase activity predicted memory outcomes more directly. Also, study-phase activity did not contribute to memory-variance that was not shared by test-phase activity. The multivariate pattern analysis also offered additional important insights. Across investigations, performance of the multivariate classifiers was positively correlated with participants' performance, and was meaningfully large for better-performing participants. This could suggest that brain activity for better-performing participants has a greater task-relevance, which is picked up by the classifiers, leading to better predictions. The multivariate pattern analysis of test-phase activity for item recognition also showed that depending on the time it takes to reach a decision, memory judgments could be driven by either a unitary, integrated signal or two independent sources of evidence; suggesting a way to reconcile the existing debate on single- versus dual-process theory. Overall, these investigations showed that predictive approaches can be used to characterize as well as to quantify the contributions of different brain-activity signals in explaining the variability in learning outcomes. While the classifier approach could be built upon to improve the classification rates, the overall modest predictions could also suggest that successful learning depends on other factors that are not reflected in the brain activity during the study- or test phases of the item recognition task, or during the feedback processing of the trial-and-error learning task. Accordingly, future investigations will need to identify and include these factors into the predictive analysis incrementally, in order to reach a more comprehensive explanation of learning behaviour.

Preface

This thesis is an original work by Sucheta Chakravarty. All research projects contributing to this work received ethics approval from a University of Alberta Research Ethics Board, project names "Organisation and Retrieval Timecourse of Human Memory," and "Organisation and Retrieval Timecourse of Human Memory across Lifespan," project numbers Pro00009760, and Pro00014801, respectively. The first study, presented in Chapter 2 of this thesis has been published as Chakravarty, S., Chen, Y. Y., & Caplan, J. B. (2020). "Predicting memory from study-related brain activity.", *Journal of Neurophysiology 124 (6)*, 2060–2075. Copyright the American Physiological Society. Reprinted with permission. I conceived and designed this research along with Yvonne Y. Chen and Jeremy B. Caplan. I also conducted the data analysis, interpreted the results, prepared the figures, wrote the manuscript and revised it. Experimental data were collected by Yvonne Y. Chen, who also contributed to data analysis, interpretation of results and edited the manuscript along with Jeremy B. Caplan. Matthew H. Danyluik contributed to the data analysis presented in Chapter 3 and Jeremy B. Caplan edited this chapter. For Chapter 4, I collected and analyzed the data, and wrote the manuscript. Isha Ober contributed to experimental design and initial collection of data and analysis. Isha Ober, Christopher R. Madan, Yvonne Y. Chen, Esther Fujiwara and Jeremy B. Caplan contributed to the conceptualization of the experiment and edited the manuscript.

Acknowledgements

I am indebted to many people, without whom this work would not have been possible. Most of all, I would like to thank my advisor, Dr. Jeremy Caplan, for his constant support, encouragement and guidance at every step of writing this thesis. I am also grateful to my supervisory committee, Dr. Anthony Singhal and Dr. Kyle Mathewson, for their helpful advice and encouragements. Thanks to my final exam committee, Dr. Tim Curran and Dr. Patrick Pilarski, for their insightful feedback. Special thanks to Dr. Yvonne Chen and Matthew Danyluik for their helpful suggestions on the chapters. I also thank my fellow lab members, Felicitas Kluger, Ryan Moukhaiber, Ulises Rodríguez Domínguez, Tamari Shalamberidze, Jeremy Thomas, and many undergraduate research assistants for their support and friendships. I would also like to acknowledge all the funding supports that made my research possible, the Natural Sciences and Engineering Research Council of Canada (NSERC), Alberta Gambling Research Institute (AGRI), and Department of Psychology of the University of Alberta. Last but not the least, I would like to thank my family and friends for supporting and encouraging me throughout my journey.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACC	Anterior Cingulate Cortex
AUC	Area Under the Curve
EEG	Electroencephalography
ERP	Event-related Potential
FRN	Feedback-related Negativity
HC	Hippocampus
LDA	Linear Discriminant Analysis
LPC	Late Positive Component
LPP	Left Parietal Positivity
ML	Machine Learning
MTL	Medial Temporal Lobe
RL	Reinforcement Learning
ROC	Receiver Operating Characteristic
RPE	Reward Prediction Error
RSE	Retrieval success effect
RT	Response time
SME	Subsequent Memory Effect
SMOTE	Synthetic Minortiy Oversampling Technique
SVM	Support Vector Machine
SW	Slow Wave

Chapter 1

General Introduction

The central idea motivating this work is that cognitive processes underlying the way we learn and remember are captured by recorded brain activity, and thus, can be used to explain the variability in learning outcomes, in other words, as functions of brain-activity signals. However, to understand behavioural outcomes as functions of brain activity, it is equally important to test if those can be predicted from brain activity. Statistically, predicting behaviour is different from explaining behaviour, and here, through a series of investigations I show that predictive approaches can offer complementary insights about behaviourally-relevant brain activity, when compared to more commonly-used explanatory- and descriptive methods.

Two different learning situations were considered; in one situation, participants studied lists of words, followed by old/new recognition judgments for the studied (targets) and new (lures) words. In the other learning situation, participants learned the stimulus-response rules for a large set of words through trial-and-error and with reward feedback. For both situations, I asked whether variability in participants' learning success could be explained by brain activity present when they were studying the information. In other words, does study-phase activity predict subsequent learning success or failure at test? This question was motivated from research using the "subsequent memory effect" framework (Sanquist, Rohrbaugh, Syndulko, & Lindsley, 1980), which has examined brain-activity signals during the study-phase, that support subsequent memory-success at test. However, brain-activity

signals identified with the subsequent memory effect framework are based on planned comparisons, which limit their interpretation as "predictive" of later memory, unlike what has been suggested by some researchers (e.g., Wagner et al., 1998).

This introductory chapter is organized as follows. First, to set the motivation for this work, I start with a brief discussion of what is meant by a prediction and the importance of testing for predictions. This is followed by a walk-through of the main investigations, including discussions of specific research questions and results. Lastly, I present a snapshot of the main methods that were followed across the investigations.

1.1 Making predictions

As explained in detail by Yarkoni and Westfall (2017), it is easy to confuse explanations with predictions. Conceptually, explaining a phenomenon should include the provision for predictions within it. However, the common statistical tools used for explanatory and predictive analysis do not assume this. Thus, models that explain well can be very different from those that predict well, because their goals are different. With predictive approaches, the goal is to forecast the outcome of future events. Thus, for problems in the field of cognitive neuroscience, given brain activity measures (X), the goal will be to predict the behavioural outcome (Y). To understand how the predictive goal is different from explaining, Shmueli (2010) notes (p. 293), "measurable data are not accurate representations of their underlying constructs." With explanatory approaches, the functional construct (f) between brain activity and behaviour is considered first. These are based on existing theories. Then, empirical data are collected to test it. Notably, existing theories on the functional construct of brain activity are less than complete, also making f less than complete.

With electroencephalographic (EEG) recordings, event-related potential (ERP) signals are used very commonly to identify brain activity specific to the processing of a stimulus, making a response, etc.; ERP signals are deflections from baseline EEG activity, often seen as peaks, and are characterised by their latency, polarity and scalp distribution. Importantly, ERPs are quantified by averaging across many trials, and the averaged signals are

used to make inferences about their involvement in specific perceptual-, cognitive- or motor processes. Further, ERP analysis follows a planned-comparisons and descriptive approach, which run the risk of overestimating the ERP features by an error commonly known as overfitting (overfitting is explained in detail in the next section). They could also underestimate what really drives the difference in the processing of the event by overlooking other relevant features of brain activity. In contrast, with predictive models, f is estimated from the empirical data, and thus, predictive models may better identify the relevant features. See the "ERPs" paragraph and Figure 1.5 below which explain why a descriptive result may not be predictive, with an example. However, one downside is that predictive models could lead to complex constructs that are difficult to interpret with existing theories.

That said, the motivation for using predictive approaches, as elegantly summarized by Shmueli (2010) is easy to understand, and a few points are of direct relevance to the current work.

- Consider that recordings of brain activity (e.g., with fMRI, EEG) typically contain thousands of features. With limited background theories, it is difficult to test which features index a mental process and how exactly they function. In other words, theoretical constructs are not well-defined to make hypotheses about the effects of the many features present in the brain-activity recordings. In contrast, with predictive models, it is easier to analyze many features. Specifically, machine learning classifiers can automatically learn the relationship between the dependent and independent variables in case of multivariate data. Thus, exploratory predictive analysis can be used to inform existing theories, and to improve them.

- Testing existing theories against predictive benchmarks is important because it estimates how well the neural measures explain the variability in the behavioural data. The amount of predictive success can also be used to directly compare between two or more brain-activity signals and their underlying cognitive processes.

- Failure to predict behaviour with a brain-activity signal is also an important finding.

The failure could be due the fact that the signal does not contain relevant information for predicting behaviour, in which case, other relevant signals could be pursued. The failure could also be due to the fact that the technique does not have enough sensitivity to capture the information from the signal, which can motivate methodological improvements to the predictive framework.

Thus, although planned-comparisons and descriptive-methods used in ERP research have identified many interesting brain-activity features, some of which have been highly-replicated across studies, to make better connections with behaviour, these should be tested with predictive approaches.

Perhaps, the two most important concepts that explain the difference between inferences drawn from planned-comparisons and predictive-tests are overfitting and generalizability. These are briefly discussed below.

1.2 Overfitting and Generalizability

Overfitting is a type of error where the model fits the data too closely, most likely by capturing the noise in it (Bishop, 2006). The problem with overfitting is that although the model produces an extremely good fit to the specific sample dataset it had operated on, it fails to generalize that fit for other samples. In other words, the model cannot provide good fits when applied to another sample dataset. This renders the model-related inferences useless for population-level data.

Consider a situation where the underlying function connecting the dependent variable (x) and the independent variable (y) is linear in nature: $y = a + bx$, where a is the intercept and b is the slope. Now consider that in one sample, the data look like that of Figure 1.1a. Clearly, this sample can be fit easily with a linear regression. The measure of goodness-of-fit, $R^2 = 1$, suggesting a perfect fit.

However, experimental data rarely look as clean as Figure 1.1a, due to sampling-error and/or measurement-error. So, a more realistic example of sample data would look like that

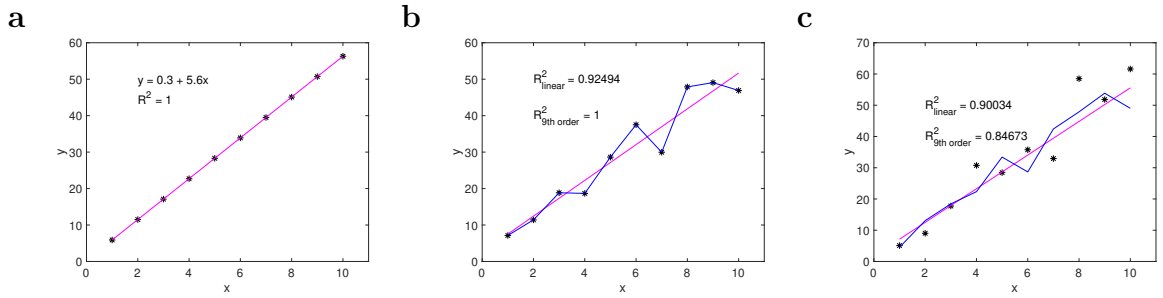

Figure 1.1: Examples of sample data and overfitting. a) A perfect sample; data contain no noise, and present a perfect linear fit. b) A more realistic sample; data contain random noise, a higher-order polynomial produces a perfect fit for the sample. c) Another sample drawn from the same population as that of panel b, but with different random noise; the fitted higher-order polynomial (from panel b) is no longer a perfect fit for the sample data in panel c.

of Figure 1.1b, where, a linear-trend between x and y still exists, but it is embedded in random-noise. Here, a polynomial-fit (9^{th}-order) looks much better than a linear-fit. The R^2 values for the 9^{th}-order polynomial-fit and the linear-fit in this case are 1 and 0.92, respectively. However, in another sample, which contains different randomly-sampled noise (Figure 1.1c), this 9^{th}-order (fitted) polynomial fits the data poorly ($R^2 = 0.80$), whereas the more parsimonious linear-fit is better ($R^2 = 0.90$). This is due to overfitting.

Importantly, for regression analysis used in the explanatory- or descriptive approach, the goodness-of-fit measures are only used to estimate how closely the model fits the sample data. Thus, unless all other samples from the population are extremely similar to the chosen sample, there is no guarantee that any higher-order polynomial-fit will generalize across different experiments. That said, a line-fit, due to its smallest degrees-of-freedom, is never likely to overfit the data. It can, however, underfit the data. Underfitting refers to the situation where the model lacks important parameters that are necessary to explain the data-variability.

When predicting behaviour from brain activity, overfitting can also be a problem for a different reason. There could be idiosyncratic reasons, and cognitive or neural processes that correlate with those reasons, for a particular behavioural outcome that do not generalize to other behavioural outcomes. For example, the word Brook may be memorable to someone

because it reminds them of their dog named Brooks; accordingly, brain activity related to the pet may *describe* why they remembered Brook in one sample dataset, but it will not help explain why they might remember Tomato in a different sample.

Unlike the explanatory methods, with predictive analysis, the importance is placed on generalizable results. Thus, additional steps are typically built into the analysis to check for overfitting. Consider that one way to know whether a model overfits the data is to test it on another sample (from the same population). Accordingly, machine learning classifiers use different samples to fit or estimate the model parameters, and to test the performance of the model. These are known as the "training-set" and the "test-set," respectively. Thus, an overfitted model is likely to perform poorly for the test-set, based on which, the model can be revised. Often, it is not easy to collect training- and test-samples separately, or the total amount of available data is small. Even in those cases, cross-validation procedures can be used to create disjoint training- and test-sets (explained later in this chapter).

Notably, the goal of finding a model that generalizes to out-of-sample data may be counter-intuitive to the goal of finding a model that efficiently captures the regularities of the in-sample data. Overfitting of a model is reflected in the model variance, which measures resilience of the model when training on different training-sets. As model variance increases, the generalizability of the model decreases. Thus, the goal is to obtain models with low model-variance.

However, a low model variance can be associated with a higher model bias. Model bias is the inherent error in the model's assumption to learn the connection between the independent and dependent variables. For example, simpler models (e.g., linear models) can have a higher model bias due to making simpler assumptions, when the data actually contain complex interactions between their parameters. When the model bias is high, the model produces many false alarms for out-of-sample data. On the other hand, non-linear models, as illustrated in the example above (Figure 1.1c), can overfit when the data are simple, and thus, perform poorly for test-sets.

Thus, the test-error of the model, which is the model error for the out-of-sample data or

the test-set, is influenced by both model bias and model variance. The actual goal is then to find a right trade-off between the model bias and model variance. In other words, to perform well on test-data, a model should have a low model bias along with a low model variance.

In general, larger data sets, containing many examples, help find models with low bias and variance. In contrast, in the descriptive approach, sample size is determined based on the desired effect size or, in some cases, it is biased by when a significant effect (e.g., $p < 0.05$) is reached, also known as "p-hacking", which is more likely to support sample-induced biases in the effects than when using a considerably large sample (Simonsohn, Nelson, & Simmons, 2014).

Thus, overall, predictive approaches can be used to better estimate behaviourally-relevant brain activity. As mentioned at the beginning of this chapter, this thesis makes detailed use of predictive approaches to investigate brain activity underlying successful learning. The main questions of interest, with respect to the two different learning situations pursued here, are discussed below.

1.3 A walk-through of the investigations

This thesis includes four studies. Chapters 2 and 3 are based on the item recognition experiment; Chapters 4 and 5 are based on the trial-and-error learning experiment. For both tasks, the investigations first looked into the predictive power of previously-identified ERP-measures relevant to successful learning. Thus, the classification rule for predicting learning success versus failure, was pre-determined, and was based on the findings from the previous planned-comparisons studies. Then, the entire dataset was considered for testing the classification rule. Thus, this analysis re-evaluated previous findings against predictive benchmarks, adding valuable insights to the interpretation of the highly-replicated ERP effects.

The tests of predictions with the previous ERP measures were followed by more data-driven, multivariate pattern analysis of brain activity. The multivariate pattern analysis has the potential to find combinations of different previously-known signals, or even to identify previously-unknown signals that are also relevant to the behavioural outcomes. Thus, the

	Item recognition	**Trial-and-error learning**
Stimuli	nouns	nouns
Set size	25	48
Learning mediated through	instruction	reward feedback
Response	old/new judgments (test)	choose or not-choose the word
Main performance measure of interest	d'	accuracy (per cycle)
Repetition of Stimuli	never	on every cycle
Lures at test	yes	no
Task engagement*	variable	higher

Table 1.1: A comparison of the item recognition and trial-and-error learning tasks studied in this dissertation, based on the different dimensions of their design. *Task engagement is on speculation basis; presumed higher for the trial-and-error learning task due to the reward feedback.

multivariate pattern analysis was capable of addressing the blind spots of the planned-comparisons approach.

Table 1.1 presents a comparison of the two tasks. In item recognition, participants were asked to study the words for old/new recognition tests that followed (see Figure 1.2), whereas in trial-and-error learning, participants learned to make correct choices on a trial-by-trial basis, in order to maximize the total rewards, which were earned at the end of the experiment (see Figure 1.3). However, in order to make the correct choices, participants had to study the words, because the response rules were stimulus-specific.

Thus, both tasks included study and test trials for the words. However, unlike item recognition, there were no lures (new items present only at test) in trial-and-error learning. As I unpack in the following sections, lures are very likely to add to the variability of the learning outcomes. Accordingly, trial-and-error learning might have had less variability in learning outcomes than item recognition.

Also, the item recognition task included a short delay (filled with a math distractor task) between the study and test phases. On the other hand, for the trial-and-error learning task, one cycle was immediately followed by the next. In fact, in the trial-and-error learning task, for a given cycle, the participants were tested on their memory for the stimulus-response rules, as learned in the previous cycle.

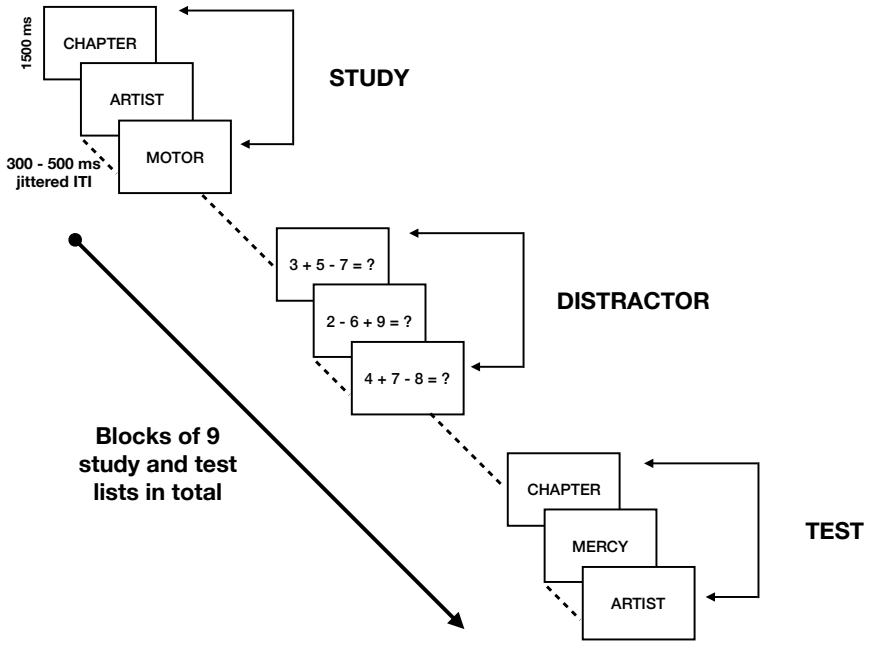

Figure 1.2: The item recognition task: participants studied lists of words, followed by a short distractor task where they solved simple math problems. After that, participants made old/new judgments for targets and lures. There were a total of 9 study and test lists. Each study and test list included 25 and 50 words, respectively. Test lists included equal number of targets and lures. Lures were never part of the study lists. None of the words were repeated for a participant. Figure is from Chakravarty et al. (2020), reprinted with permission.

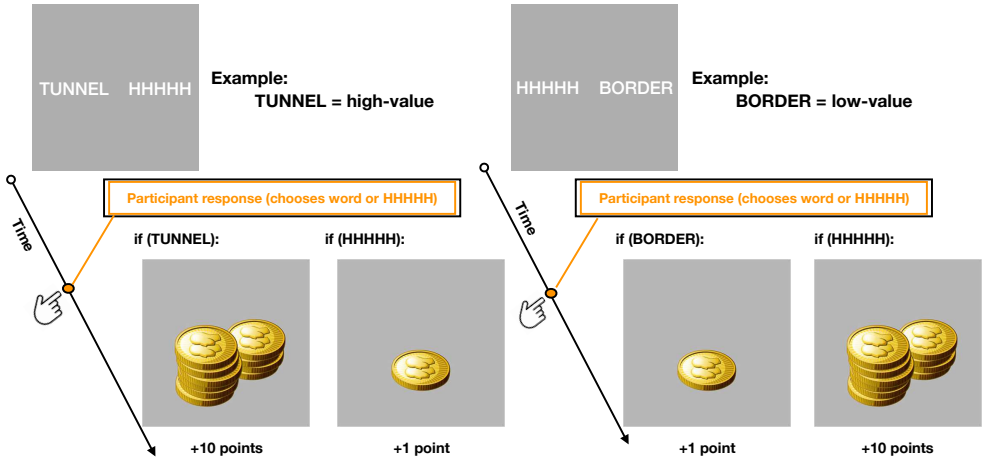

Figure 1.3: Illustration of a trial in the trial-and-error learning task. For high-value words, choosing the word led to the high (10 points) reward whereas choosing 'HHHHH' led to the low (1 point) reward. For low-value words, choosing the word led to the 1 point reward and choosing the 'HHHHH' led to the 10 points reward.

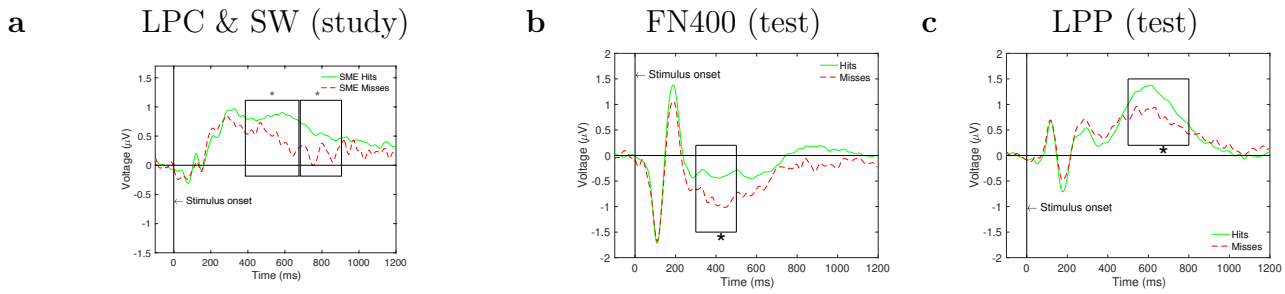

Figure 1.4: Grand averaged ERPs for the item recognition task, separated by hits and misses. a. ERPs at study, plotted for the central-parietal electrode Pz. b-c. ERPs at test, plotted for the fronto-central electrode Fz (b) and the left-parietal electrode P3 (c), respectively. The rectangles indicate the ERPs of interest: LPC and SW at study, and FN400 and LPP at test. Significant differences in average ERP amplitudes are indicated with *.

The study items (9 lists containing 25 words each) were never repeated in item recognition (neither were the lures). In contrast, the same set of 48 words were repeated across all cycles in trial-and-error learning (19 cycles in total). Years of behavioural research suggests that repeated encounters with an item leads to better probability of remembering it, though the trial-and-learning task did not include explicit memory tests for the words (but see Chakravarty et al., 2019 who investigated free recall with a similar paradigm).

The trial-and-error learning task included reward feedback, which could have led to an overall greater engagement for the task. In contrast, item recognition did not include any feedback, and thus, task-engagement likely varied across participants. In this work, multiple results from analysis of brain activity with the machine learning classifiers showed that the classifiers performed better in predicting the learning outcomes for better-performing participants. A possible reason could be that for better-performing participants, the signal-to-noise ratio (SNR) of their brain activity was higher, which was picked up by the classifiers. If this is true, then task-engagement could be an important aspect for classifier-driven analysis of brain activity.

Overall, despite the difference in the design of the two tasks, there were multiple features, based on which the investigations for the two tasks could be compared.

1.3.1 Understanding learning outcomes as functions of brain activity during the study phase

Cognitive processes present during studying of the items are thought to contribute to the variability in learning success, at the time of testing. For example, the Levels of Processing idea (Craik & Lockhart, 1972) suggests that probability of remembering an item depends on the conceptual depth with which it was studied. Accordingly, comparison of brain activity during the study phase for later remembered and forgotten items could identify signals that relate to effective encoding processes at study. Chapters 2 and 5 explored this suggestion for the item recognition and trial-and-error learning tasks, respectively.

Item recognition For item recognition, this "subsequent memory effect" (SME; Sanquist et al., 1980) has been investigated across many studies. Two commonly studied and highly-replicated ERPs at study, the late positive component (LPC) and the slow wave (SW), show more positive amplitude for subsequently remembered (hits) than forgotten items (misses) (Chen, Lithgow, Hemmerich, & Caplan, 2014; Fabiani, Karis, & Donchin, 1990; Friedman, 1990; Karis, Fabiani, & Donchin, 1984; A. S. Kim, Vallesi, Picton, & Tulving, 2009; Sanquist et al., 1980; Smith, 1993), and thus, are thought to index cognitive processes that drive the subsequent memory effect (see Figure 1.4a). Specifically, the LPC and SW are thought to index shallow- and deep encoding strategies, respectively (Karis et al., 1984). For example, when the items are repeated during the study phase, the difference due to subsequent memory for the LPC amplitude is found to be larger than the same for the SW amplitude. On the other hand, when the participants are asked to study the items using a relatively deeper study strategy, such as, generating a sentence for each item, then the difference due to subsequent memory for the SW amplitude is found to be much larger than the same for the LPC amplitude (Fabiani et al., 1990). The scalp distribution of voltage for both LPC and SW show posterior positivity. Accordingly, the centro-parietal electrode Pz is commonly used to evaluate both ERP signals. LPC and SW have different latencies; the LPC reaches its peak within the 400–700 ms time window relative to the onset of the stimulus at study.[1]. The SW is relatively sustained activity that typically takes place after 700 ms and lasts a few 100 ms long time windows, relative to the onset of the study stimulus. Thus, although LPC and SW likely supports different cognitive functions, there are similarities in the physical characteristics of the two signals; since their latencies are close and they are commonly evaluated for the same electrode Pz, it is possible that the two signals are not well separated in commonly used analysis steps. Further, as I describe in detail in Chapter 2 (Chakravarty, Chen, & Caplan, 2020), predictions for subsequent memory success for individual trials based on the amplitudes of LPC and SW produced a small significant effect only. Thus, although cognitive processes underlying LPC and SW were found to be relevant to subsequent memory

[1]The LPC is sometimes referred to as the P300 ERP signal (for a review, see Polich, 2007).

success, they did not explain a good amount of variability in subsequent memory outcomes.

As mentioned above, due to the possible overlap in their latencies and similar scalp-distribution of voltage for LPC and SW, it is possible that these signals are "process impure". In that case, multivariate pattern analysis of study-phase activity could find the right combination of the features of these signals (and features of other relevant signals present at study as well) that better predict subsequent memory success. However, the multivariate pattern analysis also produced a small significant effect (Chapter 2, Chakravarty et al., 2020). Overall, this suggests that cognitive processes present during the study phase, as reflected in the recorded (time-domain) EEG signal at study, only contribute to a small amount of variability in subsequent memory outcomes. However, the classifier performance was correlated with participants' performance, and it was meaningfully large for better-performing participants, suggesting that the chance of success of a classifier may not only depend on the cognitive processes but also how those are reflected in the recorded brain activity. Moreover, analysis of the classifier-identified pattern of study-phase activity revealed two different patterns for two different subgroups of participants in the experiment, which could suggest that there was variability in how the participants approached the study phase (Chapter 2, Chakravarty et al., 2020).

Trial-and-error learning Here, since learning was shaped by feedback, following the logic of the subsequent memory effect, cognitive processes present during feedback processing could be contributing to the variability in subsequent response accuracy (correct or incorrect choices), reflecting trial-to-trial learning. However, trial-to-trial learning is not commonly studied with the subsequent memory effect framework. Instead, it is viewed from the perspectives offered by theories of reinforcement learning, which suggest that optimal behaviour in trial-and-error learning situations can be achieved by measuring the discrepancy between the expected and actual outcomes, also known as reward-prediction error (RPE; Sutton & Barto, 1998; Holroyd & Coles, 2002). When the outcome is worse than expected, a negative RPE is generated, whereas for better than expected outcomes, the RPE is positive. RPE is

measured on a trial-by-trial basis to update the subsequent expectation as well as to update the response.

Activity of the dopamine neurons in the midbrain are found to code for RPE through phasic responses. Studies with animal models showed that when the outcome is better than predicted (or in the case of a positive RPE), there is a phasic increase in the firing rates of the dopamine neurons whereas when the outcome is worse than predicted (or in the case of negative RPE) there is a phasic decrease (from spontaneous rate) in the firing rates of dopamine neurons. When the outcome is similar to that predicted (zero RPE) no such changes in dopamine activity was observed (Schultz, Dayan, & Montague, 1997; Tobler, Fiorillo, & Schultz, 2005). In humans, there may be analogous brain regions in computing RPE, studies with fMRI show that BOLD activity levels code for RPE in the the ventral striatum (Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006) and ventral tegmental area (VTA; D'Ardenne, McClure, Nystrom, & Cohen, 2008). Further, drugs that change striatal dopamine activity are found to also change reward-related response actions (Pessiglione et al., 2006).

Previous EEG investigations of trial-and-error learning have reported an ERP, namely, the feedback-related negativity (FRN; Miltner, Braun, & Coles, 1997; Nieuwenhuis, Holroyd, Mol, & Coles, 2004), which is elicited during processing of the outcome and is likely generated from the anterior cingulate cortex (ACC) (Hauser et al., 2014; Gehring & Willoughby, 2002; Van Veen, Holroyd, Cohen, Stenger, & Carter, 2004; Miltner et al., 1997), a site that receives dopaminergic input from the midbrain regions (Rolls, McCabe, & Redoute, 2008). Accordingly, the FRN may index RPE. The FRN amplitude is more negative following unexpected than expected feedback-outcomes (Bellebaum & Daum, 2008; Hajcak, Moser, Holroyd, & Simons, 2006; Holroyd et al., 2004; Marco-Pallares, Cucurell, Münte, Strien, & Rodriguez-Fornells, 2011; Pfabigan et al., 2015). Larger (more negative) RPE (for error) is followed by response adjustments in subsequent trials. Thus, if FRN follows an RPE function, then larger (more negative) amplitudes of FRN are also likely to be followed by subsequent response adjustments. However, predictions for trial-to-trial learning based on the amplitudes

of an FRN-like signal found for this task, as investigated in detail in Chapters 4 and 5, produced a significant effect only when the previous trial was correctly responded, but not when it was incorrectly responded. Thus, in contrast to the RPE function, the FRN-like signal supported the maintenance of learned responses across successive cycles. Chapter 5 also included multivariate pattern analysis of brain activity during feedback processing to predict trial-to-trial learning, but the results were non-significant, most likely because small number of trials were available to train the classifiers.

Importantly, the above prediction results with the amplitudes of the FRN-like signal may have been foreshadowed by the investigations in Chapter 4, which looked into the elicitation and characteristics of the FRN-like signal for the trial-and-error learning task following traditional ERP analysis methods. Notably, cognitive neuroscientists have thus far stuck primarily to very simple tasks when studying the FRN and its potential role in tracking RPE. However, research in the field of reinforcement learning is far more complex and has investigated various learning situations, including both when learning depends on learning a rule (model based) and when it does not (model-free). In contrast, in cognitive neuroscientific research, studies looking into whether the FRN acts as an reinforcement learning error signal in the brain, have typically considered learning situations like the two-armed bandit (Gehring & Willoughby, 2002; Goyer, Woldorff, & Huettel, 2008; Hajcak, Moser, Holroyd, & Simons, 2007). Motivated from the situation of a gambler operating on two different slot machines, the participant bets on two different stimuli that give out rewards with different probabilities. Thus, reward-prediction itself is the main goal for a gambling task. Unlike this, in the trial-and-error learning task investigated in Chapters 4 and 5, reward-prediction was a way to learn the stimulus-specific response-rules. The results of Chapter 4 showed that although an FRN-like signal was present, it was recruited differently by the task than what would be expected from a purely RPE function. Also, the scalp distribution of voltage for this signal was more frontal than the mid-frontal negativity typically observed for the FRN. Thus, it is possible that the FRN-like signal identified for the current trial-and-error learning task was characteristically different from the FRN signal reported by previous studies. The FRN-like

signal was still relevant for trial-to-trial learning mechanisms but it was modulated by task variables that did not manipulate RPE.

Overall, the aim of evaluating learning outcomes as functions of brain activity present during the study phase, was achieved for both the item recognition and trial-and-error learning tasks. But the size of predictions were modest, which motivated exploration of other important factors, such as the test-phase activity for the item recognition task (Chapter 3), discussed below.

1.3.2 Understanding learning outcomes as functions of brain activity during the test phase

For item recognition, memory performance is measured by the ability to distinguish the targets (studied items) from the lures (new items). Since lures are only presented at test, a reasonable assumption is that cognitive processes present during test phase are more relevant to the recognition-memory outcomes than those present at study. Accordingly, Chapter 3 looked into the predictive affordance of test-phase activity. Two ERPs at test, namely, the FN400 and the late parietal positivity (LPP), may index cognitive processes that support recognition memory outcomes (Chen et al., 2014; Friedman, 1990; Neville, Kutas, Chesney, & Schmidt, 1986; Rugg & Nagy, 1989; Rugg, 1995; Rugg & Curran, 2007; Warren, 1980; Wilding & Rugg, 1996). Both FN400 and LPP show more positive voltage for hits than correct rejections, known as the old/new effect. The FN400 and LPP also show more positive voltage for hits than misses, known as the retrieval-success effect (coined by Dolcos, LaBar, & Cabeza, 2005) (see Figure 1.4b-c). Based on these previous findings, Chapter 3 investigated predictions for recognition memory outcomes, based on the amplitudes of the FN400 and LPP.

Considering the four different outcomes at test (hits, misses, false alarms and correct rejections), four classification problems were chosen:

1. Old versus New (or targets versus lures)

2. Hits versus Misses (or the same classification as that followed with the study-phase

activity)

3. Items perceived as old (hits and false alarms) versus new (misses and correct rejections)

4. Correct rejections and False alarms

Both FN400 and LPP amplitudes achieved modest success in predicting old and new trials, suggesting that these signals supported how targets and lures were discriminated in the brain. FN400 and LPP amplitudes also achieved modest success in predicting perceived-old- and new trials, suggesting that these signals were used by the participant to make the memory decisions. Classification of hits and misses was also successful, and thus, these signals likely supported difference due to memory success. However, both FN400 and LPP amplitudes failed to predict correct rejections and false alarms, which contrasted with a previous suggestion that false alarms are processed more like targets than lures, and are driven by a familiarity-based process indexed by the FN400 (Finnigan, Humphreys, Dennis, & Geffen, 2002; Wolk et al., 2006).

Notably, researchers have debated on the number of sources of evidence that drive memory judgments; a dual-process account (for a review, see Yonelinas, 2002) suggests that recognition judgments are produced by two (or more) sources of evidence independently: one is based on a sense of familiarity with the studied item and the other is based on the recall of specific details about the studied item (recollection). The FN400 and LPP are thought to index familiarity and recollection, respectively (Rugg & Curran, 2007). In contrast, a single-process account suggests that recognition judgments are driven by a unitary, integrated signal (Dunn, 2008; Wixted & Stretch, 2004). Relevant to these suggestions, multivariate pattern analysis of test-phase brain activity, as described in detail in Chapter 3, predicted different memory outcomes significantly better than amplitudes of FN400 or LPP alone, by finding better combination of the features. The multivariate pattern analysis also succeeded for the classification of correct rejections and false alarms, suggesting that there may exist memory-relevant signals beyond the FN400 and LPP.

Further, the classifiers could be used to more directly examine the dynamics of test-phase

activity prior to reaching the memory decisions. This showed that when decisions were reached relatively early (faster response times), they may have been driven by a unitary, integrated signal, which is more in line with a single-process account. However, when it took relatively longer to reach the decisions, a dual-process account became more apparent, for there were an early and a late source of evidence, which were not integrated, and drove the judgments. These results contrasted with Weidemann and Kahana (2019a) who carried out classifier analysis based on the same logic, using spectral EEG features, and found support for a single-process account only.

1.3.3 Comparison between the predictive power of study- and test-phase activity

Having investigated the predictive affordance of both study- and test-phase brain activity for the item recognition task, a natural question was how those predictions compared with each other. This was also pursued in Chapter 3. This showed that predictions based on the ERP amplitudes were not significantly different for the ERPs at study (LPC and SW) and test (FN400 and LPP). However, multivariate pattern analysis of test-phase activity predicted significantly better than both the univariate ERP measures at study (LPC and SW) and test (FN400 and LPP), as well as the multivariate pattern analysis for study-phase activity. Further, a study+test classifier, that used brain activity features from both study- and test phases, was not significantly better in predicting memory outcomes than the test-phase classifier. Together, these results suggested that memory-relevant cognitive processes at study are retrieved by those at test. In addition, cognitive processes at test bring in variability in memory-outcomes on their own (e.g., those relevant to processing of the lures).

Overall, these investigations add meaningful insights about brain activity underlying successful learning. The investigations also make the case that predictive analyses could be an important tool to advance cognitive neuroscientific theories and pave the ways for future development of important learning applications based on behaviorally-relevant brain activity.

1.4 A walk-through of the main methods

EEG recordings This work used EEG recordings, specifically, the time-domain features of EEG. These are widely used in cognitive neuroscientific research to look into electrical activity of the brain as participants perform psychological experiments. Scalp-EEG, where the electrodes are placed on the scalp, is non-invasive and extremely safe. The EEG signal measures electric potentials as a function of time. Recall that the definition of electric potential is that it is the amount of energy required to move a single unit of electric charge from one 'reference' point to a test point. Accordingly, the electric potentials measured with EEG are subject to a reference, e.g., a reference electrode. Also the raw EEG signal is very small in magnitude and has to be amplified for ease of analysis. That said, EEG does not reflect the electrical activity of individual neurons, but rather the volume-conduction of many neurons aligned similarly, e.g., the pyramidal cells. With EEG being recorded from the scalp, locating the generator of a signal has always been a challenge. Still, topographical distribution of voltage values can help understand the alignment of the neural generator (or the dipole).

ERPs One common way to analyze EEG data is with ERPs, which can isolate voltage-changes specific to the psychological events of interest. To obtain ERPs, the continuous EEG signal is time-locked into events of interest (e.g., the stimulus onset, the response etc.), and then the epoched signals for the relevant trials are averaged together. The deflections (from baseline) in this averaged signal, are thought to relate to a mixture of different cognitive processes. Importantly, ERPs are evaluated at the participant level— the ERP amplitude, which is obtained after averaging over many trials and for a specific condition A, is compared to that for another condition B, across participants.

Consider Figure 1.5a, which shows the grand-averaged subsequent memory ERPs along with the shaded error bars which are the SEM. It is clear from this figure that at the participant level, for the trial-averaged signal, subsequent hits are significantly more positive

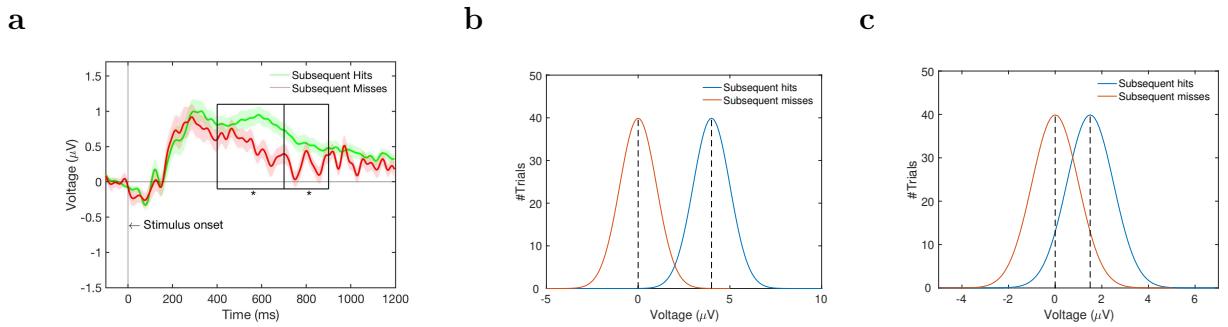

Figure 1.5: a. Grand averaged ERPs during the study phase of the item recognition task, comparing between subsequently remembered (hits) and forgotten (misses) words; ERPs are plotted for the electrode Pz. Shaded error bars represent std. error of the mean. Difference due to LPC (400–700 ms) and SW (700–900 ms) amplitude are marked with *. b-c. Distribution of the LPC amplitude across all trials (hits and misses) for two different participants showcasing that significant difference in the mean amplitudes at the participant level (a) may not necessarily imply difference between the two conditions (hits and misses) at the level of individual trials (c).

than subsequent misses, for both LPC and SW. However, if we look into the distribution of the LPC amplitude across individual trials, the situation could be different. It is possible that for one participant, the distributions of LPC amplitude for hits and misses look like that of Figure 1.5b, where not only the mean amplitudes of the two distributions are different but there is also enough separability between the distributions, so that we can set a classification threshold in the middle of the two distribution means and classify individual trials into subsequent hits and misses, expecting to be wrong in only a few cases. However, for another participant, these distributions could look like that of Figure 1.5c, where although the mean amplitudes are still different, the distributions are largely overlapping and so, a similar classification rule as above would produce wrong predictions in many cases.

Thus, ERP analysis cannot answer whether there is a good amount of overlap between the distributions of the ERP amplitudes for individual trials, for the two conditions A and B, or if those distributions are easily separable. The separability of conditions at the level of individual trials is an important consideration, for a difference in the mean amplitude along with greatly overlapping distributions is orthogonal to the suggestion that an ERP signal indexes difference between conditions A and B meaningfully. Also, the process of

averaging helps bring out parts of the signal with low trial-to-trial variability and washes out those with greater variability, while the latter can also contain information specific to the psychological events of interest. However, conducting analysis at the level of individual trials and separately for each participant could at least overcome the inter-individual differences in signal-variability. Moreover, statistical tests for the ERPs are typically conducted for a few electrodes only, selected on the basis of suggestions from previous studies; using a larger (exploratory) set of electrodes to analyze the effects renders the approach subject to tests of multiple comparisons, which is not very easy to implement. Thus, planned comparisons of ERPs also limit discovery of other relevant signals, recorded by other electrodes or a combination of the signals from different electrodes.

ROC analysis of ERP amplitudes The discussion above makes it clear as to why a descriptive result may not be predictive. To investigate if learning outcomes can be predicted from the univariate ERP measures, I followed a signal-detection theory approach (Green & Swets, 1966). Briefly, after computing the mean amplitude of the ERP for individual trials, these were sorted by magnitude (see Figure 1.6 for a demonstration). Then, a variable classification threshold was used to classify individual trials into the two conditions (e.g., subsequent hits and misses). The classification rule depended on the previously known ERP effect. So, for example, when testing for predictions for subsequent memory success based on LPC amplitudes, the classification rule was that LPC amplitude is more positive for subsequent hits than misses (see Figure 1.6). Accordingly, for each threshold value, all trials above the threshold were classified as hits, those below the threshold were classified as misses. Thus, for each threshold, there were some trials that were classified right, others were not. Based on this, the true positive- and false positive rates were calculated for each threshold. Then, the true positive- and false positive rates across all thresholds were plotted against each other to obtain the receiver operating characteristic curve (ROC). Then, area under the curve (AUC) of the ROC was computed. The AUC measured classification performance. For AUC = 0.5, classification was at chance; for AUC = 1, classification was perfect. Also,

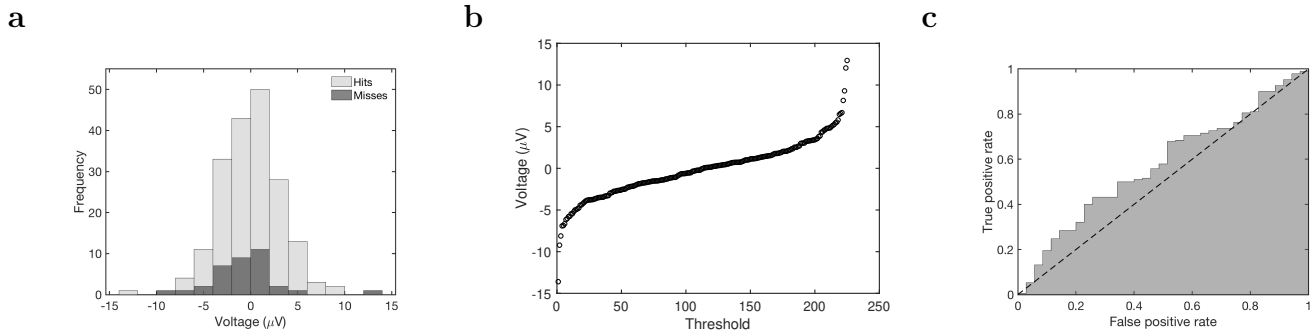

Figure 1.6: Demonstration of classification based on ERP amplitudes. a. Distribution of the ERP amplitudes across all trials for a randomly selected participant and for two different conditions, such as, hits and misses. b. The thresholds used for classification. c. The ROC curve, shaded region represents the AUC. Dashed black line denotes chance.

in this case, AUC < 0.5, suggested that evidence for an opposite classification rule than that based on previous ERP findings. So, for the LPC example above, AUC < 0.5 would suggest that subsequent misses were more positive than subsequent hits. Two-tailed t-tests were used to test if the AUCs across all participants were significantly different from chance.

Multivariate pattern analysis Importantly, the above ROC analysis of ERP amplitudes for individual trials was based on effects known from previous planned-comparisons of the ERPs. Thus, although it provides an objective measure for the separability of two conditions at individual trials level, it is possible that differences due to learning success are better measured in terms of patterns of brain activity, rather than the univariate ERP measures.

In recent years, there has been a growing interest in employing machine learning classifiers for analysis of brain activity. For example, brain computer interfaces (BCI) aim to connect brain-activity signals, measured with scalp-EEG or electrodes placed inside the brain, to an external computer. BCI research has seen many applications of the classifier methods to better identify brain activity relevant to the sensory- or motor events of interest. Classifiers are also being used in basic cognitive neuroscientific research with EEG.

However, the underlying research question may be of special consideration here. Classifier analyses are often used as "proof-of-principle" for whether or not there is behaviourally-relevant information in the brain. Although a valid question, it may not always be an

interesting one. For example, decision making is dependent on the availability of relevant information. Accordingly, brain activity at the time of decision making should include processing of that information. Thus, it may not be surprising to find that a classifier succeeds in predicting the decision, after being trained on brain activity present during making the decisions. A more interesting question here could be what are the characteristic signals that support the decision making. Another important aspect of classifier analysis of brain activity is the size of the prediction, across a decent number of participants. The size of prediction gives us clues about the amount of variability in the behavioural outcome that could be explained by the chosen brain-activity features, an important point to consider when attempting to explain behaviour as functions of brain activity.

As discussed in the walk-through of the investigations, here, I used machine learning classifiers to examine patterns of brain activity that supported successful learning. The investigations were not directed towards a proof-of-principle for information being present in the brain, but to draw inferences about behaviourally-relevant brain activity and their underlying cognitive processes. A few key points of the analysis using classifiers are described below.

Supervised classification With classification methods, we can predict discrete class labels, such as, different psychological events. This is different from predicting continuous output values, which is obtained with regression methods. The type of classification method used in this thesis is commonly known as supervised classification. Here, the model is supplied with both the psychological event-labels and the data, in order to learn the underlying regularities, which is known as training. In the other type of classification, known as unsupervised classification, the model is not supplied with the event-labels, and instead, it attempts to figure out those different classes by looking into the feature characteristics of the sample data; an example of unsupervised classification is cluster analysis. Here, with the supervised classification, the goal was to find a transformation of the high-dimensional EEG data onto a single dimensional decision space. The decision boundaries separated the

different classes, which represent different psychological events. The decision boundary can be linear or non-linear, based on which, the classifiers are also broadly categorized into linear and non-linear models.

Choice of classifiers Here, I have used two, arguably the most simple, linear models: linear discriminant analysis (LDA; Fisher, 1936) and support vector machine (SVM, with a linear kernel; Cortes & Vapnik, 1995). LDA works by optimizing the weights to the features of the data in a way so that the variance within each class is at minimum, whereas the variance between the two classes is at maximum. Thus, LDA weights can be directly translated into the relative importance of the different features of the data for the classification. In this thesis, I have frequently used the LDA weights to gain insights about the classifier-identified pattern of activity. On the other hand, SVM works to find a hyperplane with the greatest margin that separates the two classes.

Workflow of the classifier analysis Note that for both linear- and non-linear classifiers the main workflow of the analysis is generally the same. First, the model sets default starting values for its parameters. Next, when supplied with the training data, it estimates the best fitting parameter values, through optimization techniques that aims to find the minimum error, also known as the training error. Now, as with any model-fitting attempt, there is always a possibility of finding parameter estimates that correspond to a local minimum rather than the desired global minimum. To offset this, iterative steps are taken, in each of which, the starting parameter values are changed. Once trained, the fitted-model is applied to a test set to make predictions. Figure 1.7 presents a schematic overview of the main steps involved in the classifier analysis.

Cross-validation Notably, while a small value of the training error is used as a sanity check, it is of little consequence to the model performance because model performance is evaluated for the test-set. As mentioned before, classification analysis regularly uses cross-validation techniques, which separates the data into disjoint training and test sets. In the

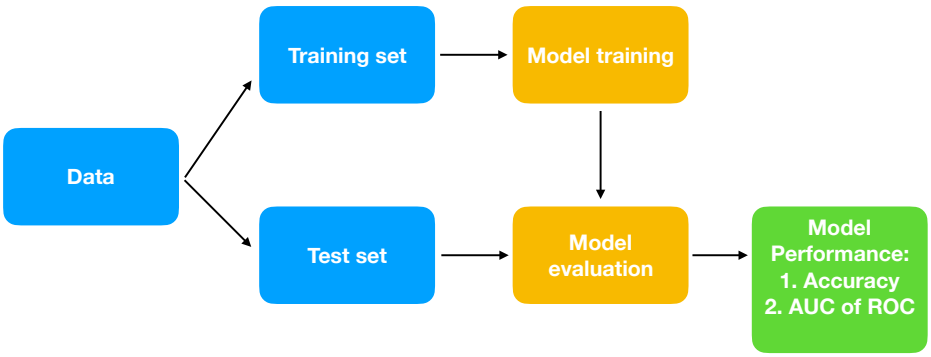

Figure 1.7: A schematic overview of the classifier analysis: data were split into training- and test sets through k-fold cross-validation, the training sets were used to train the model, while the test sets were used to evaluate them; the final model performance was averaged across all the test sets.

current work, I used k-fold cross-validation. Here, the data are randomly partitioned into k-folds, where k is an integer. Then, $(k-1)$ folds are used to train the model and the remaining fold is used to test it. This process is repeated k times, so that each fold is used as a test-fold exactly once. The overall performance of the model is usually measured by its average performance across all the test-folds.

Curse of dimensionality Notably, as in many other fields, such as image-processing, classifying EEG data is challenged by the greater number of features available, in comparison to the very small number of samples, e.g., number of trials available for within-subject classification. As mentioned before, with more features than observations, chance of overfitting increases. To understand this situation, consider again the example of the regression I had presented before (Figure 1.1b); the 9^{th} order polynomial had almost as many parameters as the number of observations (10) and thus it could assign one parameter per observation. This is opposite to the idea of finding regularities in the data; instead it aims to "catch" the given observations only, and it is possible to do so because of the greater number of available parameters. Thus, a fit obtained this way is bound to fail even for small changes in the observations.

Feature selection Likewise, with EEG, usually there are thousands of features, whereas the number of samples is at most a few hundreds. To handle this, researchers usually select a subset of features, either based on prior knowledge or based on feature-selection or feature-extraction methods, that are included within the classification analysis. Feature reduction methods, which transform the high-dimensional EEG data to a relatively low-dimensional space are also used in some cases. Importantly, for any method that uses the class-information to select, extract or to reduce features is subject to a circular logic if applied to the whole dataset. Accordingly, caution must be taken to inform these methods only based on the class-information in the training data. In this thesis, I pre-selected features based on general knowledge of EEG recordings. For example, to reduce the spatial features,

a small subset of 10 electrodes were selected so as to roughly cover the scalp. Also, the signal from each selected electrode was binned over 100 ms time-windows to reduce the number of temporal features.

Time-domain EEG features Notably, the time-domain features of the EEG signal, as used in this thesis, have one notable limitation; the latency of a signal, which is the time when it peaks, can fluctuate from one trial to another, potentially creating a challenge for the classification analysis based on individual trials. Accordingly, researchers taking related approaches have opted for the spectro-temporal features of the EEG instead, where this is less of a problem due to averaging over a greater time-window (e.g., Weidemann & Kahana, 2019a). However, it is also important to understand the applicability of the classifier methods with the time-domain EEG features, for both basic knowledge as well as due to the reason that such features could provide better time-specific information (e.g., Noh, Liao, Mollison, Curran, & de Sa, 2018).

Variability in the classifier methods Published classifier methods in cognitive neuroscientific research seem to vary from one study to another. The methods vary even for studies from the same research laboratory. There may be independently justifiable reasons for many of those choices, and it could be specifically targeted to address smaller classification rates. For example, the choice a linear or a non-linear classifier varies across studies, and till date there is no consensus about which is better for analyzing EEG data. Likewise, studies have also differed based on their choice of the specific classifier algorithm, e.g., LDA, SVM, decision trees, convolutional neural networks, etc.. Different from this, here I took the following approach:

1. use the simplest justifiable classifier methods

2. keep the classifier methods as consistent across studies as possible

The choice of linear classifiers for the current studies was based on the argument that for non-linear methods, the models generally try to capture both the individual effects and

the interactions between the features, which may be limited by the small number of trials available to learn these characteristics and will risk overfitting it more than linear models. Also, EEG voltage values tend to have a Gaussian distribution, and thus, are well suited for classifiers like LDA.

Regularization To further reduce the chance of overfitting, for the current models, I used regularization, which restricts the model from becoming too complex. Recall that the classifier uses optimization techniques to minimize the training error (or training loss). With regularization, we add a term to this loss function, which better prevents the algorithm from converging to local solutions, and scales the model parameters so as to penalize those parameters with very large estimated values. The specific regularization steps for the models are explained in the methods section of Chapter 2 (Chakravarty et al., 2020); Chapters 3 and 5 also used the same steps.

Performance of the classifier Unlike the choice of the classifier, the choice of measure to evaluate the performance of the model can be determined more objectively. For models trained on comparable number of examples from each class, model performance can be evaluated by calculating its accuracy, which is defined as, $\frac{(TP+TN)}{(P+N)}$, where TP and TN refer to the true positives and true negatives respectively. However, in situations where one of the classes is over-represented, accuracy may not provide a true estimate of the model performance, because the model may be biased towards predicting the over-represented class, and thus, makes a lot of hits or true positives but also a lot of false alarms. Consider a disease that is contracted by 1 out of 100 people only. A classifier that is set up to predict the disease, will have near-perfect accuracy if it predicted "no disease" for all cases. However, it will also have a 100% false alarm rate (or false negative rate), because it will miss all positive cases.

Since class-imbalance was present for multiple classification problems considered in this work (e.g., when classifying hits and misses for the item recognition task), I used AUC of the ROC to measure model performance. As mentioned before, ROC is obtained by plotting

the true-positive-rate against the false-positive-rate for varying classification thresholds, and AUC is the area between the two axes and the the ROC curve. For $AUC = 0.5$, the model predictions are random, for $AUC = 1$, the model performs perfectly. Models were tested on the test sets; for each observation in a test set the model produced a score, which can be transformed to its posterior class probability for belonging to one class over another. These scores across all observations in the test set were used to obtain the ROC and the AUC of the ROC. The AUCs across all test sets were averaged together.

Class imbalance In general, for imbalanced classes, to better train the models, researchers have used different approaches, including re-sampling of the trials, e.g., undersampling the over-represented class. Here, due to small number of trials overall, I used oversampling of the under-represented class instead. The oversampling was done for the training sets only and was done following the Synthetic Minority Oversampling technique (SMOTE; Chawla, Bowyer, Hall, & Kegelmeyer, 2002). SMOTE created new examples from the existing under-represented class examples. To create a new example, the algorithm 1) randomly selected an existing example from the under-represented class, 2) randomly selected one example from its k-nearest neighbours (from the same class), 3) calculated the distance between the two chosen examples, 4) added a random number between 0 and 1 to this distance and 5) added the distance (with added random noise) to the first chosen example. The new example, created this way, was in between the original example and its chosen neighbour. All of these methods were the same across the studies, and are explained in detail in Chapter 2 (Chakravarty et al., 2020).

Over-optimistic results It is also important to note that despite the use of cross-validation techniques, classifier analysis can be subject to overfitting at a higher level, producing over-optimistic results. To avoid this, it may be important to adhere to a pre-planned protocol for the classifier analysis. For example, in published work, often the reason behind choosing a specific classifier is not clearly stated; it is possible that the analysis was done using

several classifier models and only the one with the best performance was reported. Another example is the amount of regularization for the classifier model, which could have been set by evaluating the performance of the model for the test set. In both of these situations, the hyper-parameters involved in the classifier analysis are influenced by the performance of the model for the test set, when the objective behind keeping a separate test set is to use it to only evaluate and not inform the model. This problem may be similar to the "double-dipping" problem, or circular logic in analyzing brain activity. However, as explained by Skocik, Collins, Callahan-Flintoft, Bowman, and Wyble (2016), it is more difficult to identify over-optimistic results in published reports, because specific reasons behind those choices of hyper-parameters are frequently missing.

1.5 Concluding remarks and chapter overview

In sum, predictive approaches could be used to ask important cognitive neuroscientific questions, some of which cannot be addressed with planned-comparisons and descriptive methods. For example, unlike planned comparisons, with multivariate pattern analysis, it is possible to identify combinations of behaviourally relevant signals. The classifiers are strong data-driven techniques and thus, they do not depend on previous theories.

The strength of the classifiers can be appreciated even more for a few investigations where the classifiers were able to discriminate between psychological events in absence of differences in the behavioural report. For example, a study by Haynes and Rees (2005) found that with multivariate pattern analysis of BOLD activity for area V1, it was possible to predict the orientation of masked stimuli, even when it was not indicated in the behavioural response.

However, classifiers are also prone to circular analysis, which produces impressive effects but due to erroneous logic. These could lead to predictions that are based on information other than task-relevant processes; this happens if 1) the features are not selected carefully or 2) the classification problem itself is confounded. Accordingly, classifier-based investigations require a better understanding of those potential problems, in order to properly implement these methods and to gain valuable knowledge from them.

The goal of this thesis is to investigate characteristics of brain activity that can predict learning outcomes. This is not only important for advancing basic theories of how we learn and remember, but it also paves the way for future applications that are useful for improving learning abilities, as well as to evaluate learning effects in absence of overt behaviour. The work presented here follows an incremental approach towards applying classifiers for obtaining insights about behavioural relevance of brain activity, and for two different learning situations. Also, this work is more focused on predictions that work across a relatively large number of participants. Going beyond the "proof-of-principle" of the predictive analysis working, the investigations make careful note of the overall size of predictions, and the implication of those size of predictions in supporting the idea that learning success can be explained as functions of brain activity.

Overview of the chapters Chapters 2 and 3 present investigations for the item recognition experiment. Chapter 2 (Chakravarty et al., 2020) looks into brain activity during the study phase, in order to predict difference due to memory outcomes at test. It starts with the predictions based on individual ERP amplitudes at study, such as, LPC and SW. This is followed by multivariate pattern analysis of the study-phase activity. Chapter 3 follows a similar setup as Chapter 2, for investigating test-phase activity. First, it goes over predictions based on amplitudes of FN400 and LPP, followed by multivariate pattern analysis of test-phase activity. Additionally, chapter 3 uses the classifiers to investigate the dynamics of brain activity leading up to the memory decisions at test. Chapter 3 also includes comparison between study- and test-phase measures of brain activity, based on their size of predictions.

Chapters 4 and 5 investigates brain activity for the trial-and-error learning experiment. Chapter 4 presents analysis of the FRN, following traditional ERP methods. Specifically, it asks if an FRN-like signal is present for the task, and if it supports an RPE function. Finally, chapter 5 investigates if the FRN-like signal drives learning following the RPE function. Similar to chapters 2 and 3, the investigations in Chapter 5 start with predictions

for subsequent response-accuracy based on the amplitude of the FRN-like signal. This is followed by predictions with multivariate activity present during feedback processing.

Lastly, chapter 6 presents a general discussion of the findings, how those connect with previous research, and future directions. Chapter 6 also presents a critical estimation of using classifiers as the main approach for this work.

Chapter 2

Predicting subsequent memory from brain activity during the study phase of item recognition

Abstract

To isolate brain activity that may reflect effective cognitive processes during the study phase of a memory task, cognitive neuroscientists commonly contrast brain activity during study of later-remembered versus later-forgotten items. This "subsequent memory effect" method has been described as identifying brain activity "predictive" of memory outcomes. However, the modern field of machine learning distinguishes between descriptive analysis, subject to overfitting, and true prediction, that can classify untrained data. First, we tested whether classic event-related potential signals were, in fact, predictive of later old/new recognition memory (N=62, 225 items/participant); this produced significant, but small predictive success. Next, pattern classification of the multivariate spatio-temporal features of the single-trial EEG waveform also succeeded in predicting memory. However, the prediction was still small in magnitude. In addition, topographic maps suggested individual differences in sources of predictive activity. These findings suggest that on average, brain activity, measured by EEG, during the study phase is only marginally "predictive" of subsequent memory. It is possible that this predictive approach will succeed better when other experimental factors known to influence memory outcomes are also integrated into the models.

2.1 Introduction

To analyze brain activity underlying successful memory formation, cognitive neuroscientists have adopted the so-called "subsequent memory effect" (SME; Sanquist et al., 1980), contrasting brain activity during the study phase of a task for subsequently remembered (hits) versus forgotten (misses) items. The SME is a major advance over prior methods that compared activity between different encoding conditions rather than relating it to eventual memory outcomes (for reviews of the SME, see Wagner, Koutstaal, & Schacter, 1999; Paller & Wagner, 2002; H. Kim, 2011). The SME approach has produced several highly replicated findings, including the late positive component (LPC) and the slow wave (SW) of the event related potential (ERP) of the EEG, both more positive for subsequent hits than misses (e.g., Sanquist et al., 1980; Karis et al., 1984; Chen et al., 2014; Fabiani et al., 1990; A. S. Kim et al., 2009; Friedman, 1990; Smith, 1993). The robustness of the SME could be due to the fact that it indexes brain activity which is coupled with behaviour. For example, the Levels of Processing concept (Craik & Lockhart, 1972) holds that the likelihood of an item being remembered depends on the conceptual depth with which the participant evaluates or interacts with the item during encoding. Evidence has suggested different SME ERPs reflect these different processing-levels (Sanquist et al., 1980; Paller, Kutas, & Mayes, 1987; Fabiani et al., 1990).

Notably, the SME is often described as identifying brain activity "predictive" of memory success (Brewer, Zhao, Desmond, Glover, & Gabrieli, 1998; Wagner et al., 1998). If truly predictive, the SME could form the basis of important learning applications, such as tracking learning progress or testing the effectiveness of different training protocols (Fukuda & Woodman, 2015; Arora et al., 2018). Here, we examine whether "predictive" is an accurate characterization of SME ERP signals. Consider that in the traditional approach, SME ERPs are analyzed by the difference in brain activity at study for subsequent hits and misses, averaged across many trials. The hits–misses contrast is tested for statistical significance, with participants as repeated-measures. This captures the association between study-phase brain activity and memory outcomes, but such a descriptive model is not aimed at explaining the causal relationship or making predictions about new observations. To evaluate whether or not memory outcomes can be predicted from the SME ERPs, it is important to apply predictive models, which remain under-explored in this context. Predictive models could

be particularly helpful in bridging the gap between the existing theories and the potential learning applications.

Some recent studies suggest that prediction of memory from study activity could succeed with fMRI (Watanabe et al., 2011; Lee, Brodersen, & Rudebeck, 2013) as well as with intracranial EEG recordings (Weidemann et al., 2019; Weidemann & Kahana, 2019b; Ezzyat et al., 2017; Arora et al., 2018). Also, with standard, scalp-recorded EEG, Fukuda and Woodman (2015) showed that two pre-identified SME EEG measures, amplitude of the frontal slow wave and occipital alpha-band power, could predict the old/new confidence ratings given by the participants at test. Although this is a valuable finding, it skips predicting the memory outcome itself, our current aim. To classify subsequent memory from *multivariate* EEG activity at study, Noh, Herzmann, Curran, and de Sa (2014) used two classifiers. One was trained with pre-stimulus spectral features. The other used during-stimulus features, and was further divided into a time-domain and a spectral-domain classifier. Overall classification accuracy was near 60%, but the authors did not report the success rate, alone, of the time-domain signal during-stimulus. On the other hand, using only time-domain features (pre- and post-stimulus onset), Sun et al. (2016) found success in predicting subsequent memory for a majority of their participants ($N = 9$) with a convolutional neural network classifier [mean accuracy: 72.07%], whereas linear classifiers failed to classify the majority. Unlike linear classifiers, non-linear classifiers (such as convolutional neural networks) can evaluate the interactions between features, which could have led to this difference in success rates. Notably, the number of examples available to build or train these models in this type of context is usually very small for such interactions to be captured reliably. Also, it is possible that in some cases, the authors may have tried various variations of the classifier and for various reasons (including length limitations), only reported the best outcomes. This could result in inflated apparent success rates of classifiers (Skocik et al., 2016). In sum, for scalp-recorded EEG, it remains unknown if highly replicated ERP SME features are predictive. This is an important step in estimating a "benchmark" of predictive strength for this purpose. It is also unclear if the multivariate time-domain EEG signal can be used predictively. One possible limitation of the time-domain features is sensitivity to trial-to-trial variability in latency (Luck, 2014), which could impact classifier training. Thus, it is possible that researchers have tried and failed in the time domain, and opted to stop pursuing this goal in favour of spectral features (i.e., a file-drawer problem).

In the present study, we seek to understand the general prospect of using study related EEG time-domain features to predict memory with the help of easy to interpret, linear predictive models. First, using concepts from signal-detection theory (Green & Swets, 1966), we ask if it is possible to predict memory outcome (i.e., hit or miss) for each study item based on individual, previously identified SME ERPs (mean amplitude of the LPC or the SW). We consider the probability distribution of the SME ERP for all the hits versus misses and test for the amount of separability between these two distributions which can support predictions for individual trials. The predictions are made based on the rule that hits are more positive than misses (for LPC or SW, as per prior findings) and by varying the classification threshold across the two distributions.

Importantly, the SME ERPs were identified through trial-averaging and planned comparisons, which could limit their use to predict memory. Trial-averages in the traditional, descriptive analysis can help raise the signal-to-noise ratio (SNR). But while this step can identify portions of the signal with low variability across trials, it can also wash out components with greater variability, which could also carry meaningful information related to memory-encoding. With planned comparisons, the electrodes of interest are based on prior studies, thus, possibly missing out on other relevant sources of activity. To move beyond these limitations, our next step was to use multivariate measures that include features beyond those known from the traditional research. The multivariate features were then analyzed with predictive models borrowed from the machine-learning literature. These models can automatically learn useful patterns from multivariate measures (Norman, Polyn, Detre, & Haxby, 2006) and are trained and tested on separate sets of data to evaluate its generalizability; a practice that checks for over-estimation of a model and is not, in general, looked at in descriptive analyses. Thus, with this approach, we can ask the more general question: does brain activity during the study phase predict memory at the test phase?

Another motivation for comparing the univariate predictors with multivariate, machine-learning classifiers was that the standard old/new recognition task is, arguably, impoverished. These kind of judgements are highly likely to be driven by more than one process, which are reflected in more than one neural activity measure. The multivariate classifiers are designed precisely for problems such as this (multiple predictors). These have the potential to discover multiple processes, and produce a combined prediction based upon this process-impure signal. Note that it is also possible to request additional subjective judgements,

such as remember/know distinctions or confidence levels, to isolate multiple processes that are thought to underlie memory retrieval, such as recollection versus familiarity. This has been frequently done in previous classifier approaches to study phase as well as test-phase activity (Noh et al., 2014; Fukuda & Woodman, 2015; Noh et al., 2018; Liao, Mollison, Curran, & de Sa, 2018; Sun et al., 2016). However, there is the risk that this approach may alter the way participants approach the task (Eldridge, Sarfatti, & Knowlton, 2002; Hicks & Marsh, 1999). Also, this relies on subjects' ability to cleanly separate their own familiarity versus recollection processes (see Dunn, 2008, whose the state-trace analysis casts a doubt on such ability). Moreover, the issue of process impurity likely goes far beyond the recollection/familiarity distinction. Thus, to avoid this, our task included simple instructions for the participants (simply to study for a later memory test) and a simple response (old versus new). This also welcomed subject variability that could reveal interesting individual differences.

Regarding the recollection/familiarity distinction, specifically, the common dual-process view of ERPs in recognition-memory paradigms is that the FN400, an ERP elicited during the test phase of the task, reflects familiarity-based retrieval and the Late Parietal Positivity (LPP), also elicited during the test phase, reflects recollection-based retrieval (e.g., Rugg & Curran, 2007). In our data set, both of these signals produced significant old/new effects (see Chen et al., 2014), suggesting that both familiarity and recollection processes appeared at test. This confirms the process-impurity of the task. However, the FN400 (contrasting hits versus misses) covaried significantly with performance (d' and negatively with response time) across participants, whereas the LPP did not. This suggests that the putative recollection process, although clearly present, played a far more minor role in driving the old/new judgement than the putative familiarity (or conceptual priming; Voss & Paller, 2009) process. This led to clear predictions for the current classifier approach. Because the LPC SME covaried significantly across participants (Chen et al., 2014), with both the FN400 and performance (d' and response time), we expected that the LPC would produce above-chance classification of subsequent hits versus misses. Because the SW SME covaried significantly across participants with the LPP, but not with performance measures, we expected that the SW would not be able to classify subsequent old/new recognition above chance. Alternatively, variability reflected by the SW might be unrelated to individual differences, but could still support classification above chance when attempted within-subjects, as we do here.

2.2 Materials and Methods

2.2.1 Behavioural materials and procedure

Data were from the 64 participants for whom the traditional analysis of ERPs was previously reported in Chen et al. (2014). Of these, two participants were excluded for having more than 15% of the total number of study trials (225) rejected due to artifacts. Participants provided written, informed consent for the procedures. The research was approved by a University of Alberta ethical review board.

The experiment involved alternating study and test phases (Figure 2.1). Participants were given a very simple instruction— to study the words for later tests. No instructions related to study strategies were provided. For each study list, they were instructed that they will see 25 words that are to be studied. Participants were not required to make any response during study. For each test list, they were asked to make old/new judgments by pressing the relevant key (1 for old, 2 for new). Words were presented one at a time, both at study and at test. Each study word was displayed on the screen for 1500 ms with a jittered inter trial interval (300–500 ms). Each study list consisted of 25 words and was followed by a short math distractor task, consisting of 5 addition or subtraction problems involving integers from 1 to 9. The math problem remained on the screen until the participant made a response. Each test list immediately followed the math distractor task and consisted of 50 words, 25 of which were from the study (i.e., "old") and 25 were lures or "new" words. Each test word remained on the screen until the participant pressed either key. Hits were correctly responded study trials and misses were incorrectly responded study trials.

2.2.2 EEG methods

EEG was recorded in an electrically shielded, sound-attenuated chamber, from high-density 256-channel Geodesic Sensor nets (Electrical Geodesics Inc., Eugene, OR). Signal was amplified at a gain of 1000 and was sampled at 250 Hz (impedance below 50 $k\Omega$ and referenced to the vertex electrode, Cz). EEG signal was pre-processed with the EEGLAB toolbox (`http://sccn.ucsd.edu/eeglab`; Delorme & Makeig, 2004), running in MATLAB. It was bandpass filtered to 0.5–30 Hz and average re-referenced. Independent component analysis (ICA) was used to look for artifacts in the signal (such as eye blinks, muscle noise etc.). EEG trials were then epoched from 100 ms pre-stimulus to 1200 ms post-stimulus intervals.

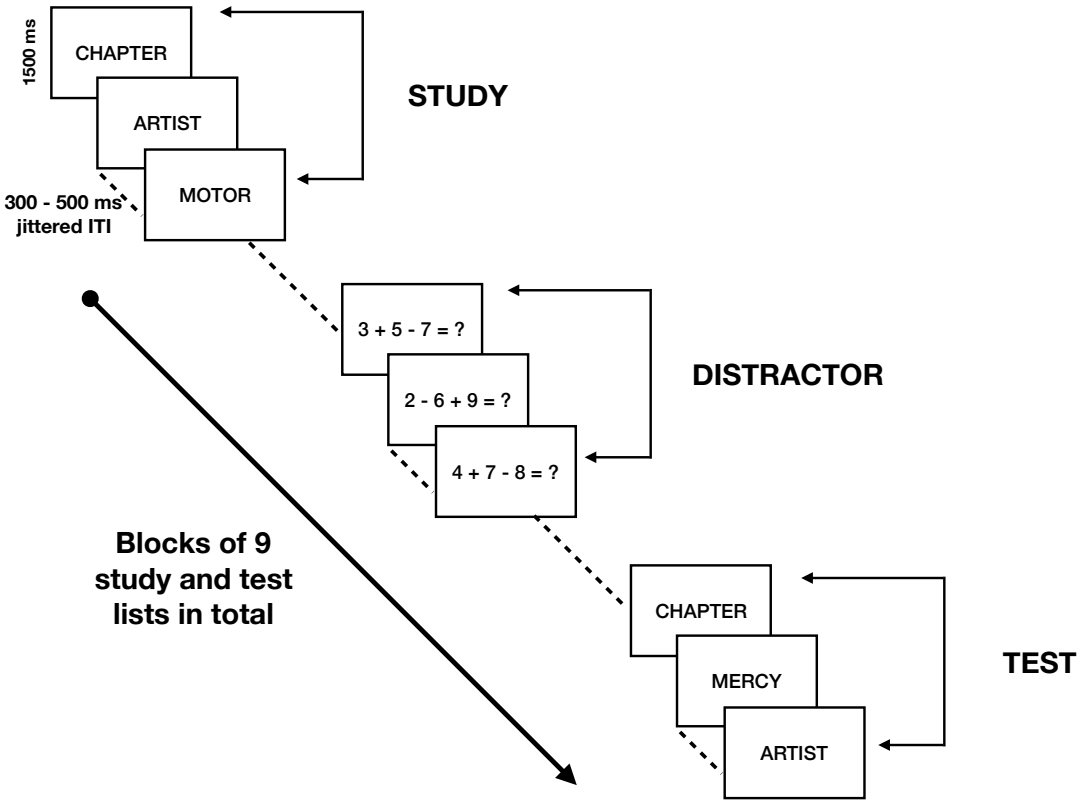

Figure 2.1: The experimental paradigm. Participants were asked to study a list of 25 words, presented one at a time at the center of the screen. This was followed by a short distractor task with simple math problems. Participants were then given a set of item recognition tests, judging each word as "old" (targets) or "new" (lures). There were equal number of targets and lures in the test phase. This whole process was repeated 9 times, yielding 225 study and 450 test trials. Each study list was unique. The order of the items during study was same as the order of the targets at test, with lures being presented at random positions in the list; lure items were not repeated across lists nor within lists.

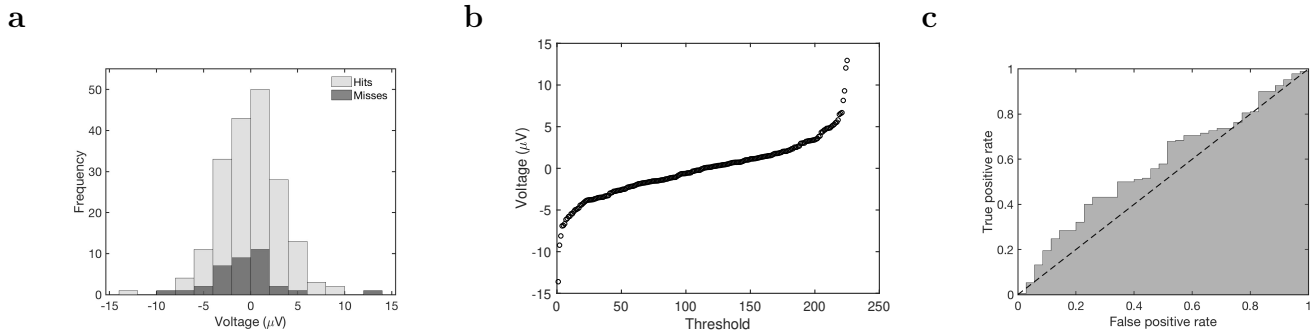

Figure 2.2: Demonstration of classification based on SME ERPs. a. Distribution of the LPC amplitude (from Pz) across all trials for a randomly selected participant. b. The thresholds used for classification. c. The ROC curve, shaded region represents the AUC. Dashed black line denotes chance.

After removing the baseline, we used a voltage threshold of 50 μV to remove epochs with large drifts. Additionally, for each epoch, we calculated the difference in voltage between adjacent time samples or the point-to-point difference to detect artifacts. We rejected epochs for which the point-to-point difference exceeded 25 μV. With the voltage and point-to-point difference thresholds in place, more than 15% of epochs were rejected for two participants (16% and 43% epochs rejected, respectively). Data from these two participants were excluded. For all other participants included in this study ($N = 62$), on average, 2 out of 225 epochs were rejected [min = 0, max = 23].

2.2.3 EEG Classification

We seek a function f that can predict discrete class labels Y (hit or miss in this case) to each trial X, i.e., $f(X) = Y$. X is a $N \times T$ matrix where N denotes the number of electrodes and T denotes the number of voltage samples as a function of time; $N \in \mathbb{Z}$ and $T \in \mathbb{R}$. Elements of X are called "features." Thus, f transforms the high dimensional space of EEG features to a one-dimensional decision space. f is called a classifier. First, we tested if classic SME ERPs, such as the LPC or the SW, when computed for individual trials, are able to predict memory outcome for individual trials better than chance. Next, we tested if multivariate pattern analysis of EEG trials from the study phase can predict memory outcomes.

Classification based on SME ERPs

Two study-related ERPs were considered, consistent with prior research dating back to Karis et al. (1984): the LPC and the SW, from the centro-parietal electrode Pz. LPC is positive going, occurs between 400–700 ms after stimulus onset and more positive for hits than misses. SW is relatively sustained activity, occurring between 700–1200 ms. Across different SME studies, SW is reported for both centro-parietal and frontal electrodes. But frontal SW is thought to reflect item–item associations (A. S. Kim et al., 2009) or processing of emotional stimuli (Diedrich, Naumann, Maier, Becker, & Bartussek, 1997; Simon-Thomas, Role, & Knight, 2005). Because we used isolated common nouns, we did not expect to see the frontal SW. The SW was subdivided into an early (700–900 ms post-stimulus) and a late (900–1,200 ms) component (see Chen et al., 2014).

For each SME ERP and for each study trial, we calculated the mean amplitude from electrode Pz, over the respective time window. The classification rule or function, based on prior (descriptive) SME results, was that subsequent-hits should have more positive voltage than subsequent-misses (Chen et al., 2014; Karis et al., 1984). Then, the receiver operating characteristic (ROC) curve was traced by setting each observed mean amplitude value as a classification threshold and plotting true positives (subsequent hits that were greater than or equal to the threshold) against false positives (subsequent misses that were greater than or equal to the threshold). After obtaining the ROC, the area under the curve (AUC) of the ROC was calculated through trapezoidal numerical integration implemented by the perfcurve function in MATLAB R2018a (see Figure 2.2 for a demonstration). AUC indexed the capability of the classifier to make more hits and less false alarms. AUC = 0.5 would reflect random predictions (chance), and a perfect classifier would achieve AUC= 1. Also, in this case, AUC$<$ 0.5 would indicate that subsequent misses were on average more positive than subsequent hits.

Multivariate classification

Here, we used multiple EEG features per trial, with the speculation that other study related EEG features (beyond the known SME measures) could also be informative for making memory predictions. Each EEG epoch had 257 electrodes, sampled at 250 Hz for 1200 ms, thus there were over 80,000 features per trial. For computational simplicity, we selected a

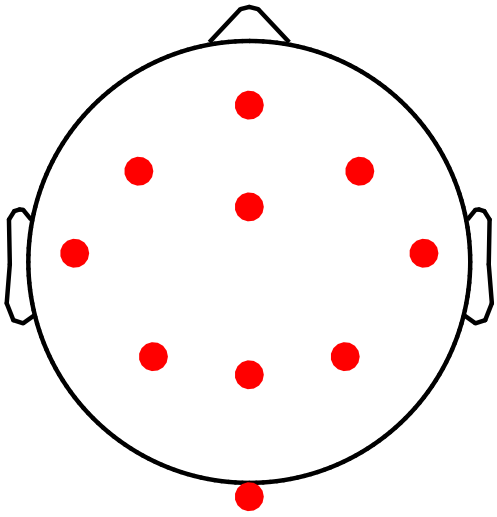

Figure 2.3: Selected electrodes for the multivariate classification, roughly distributed in equal between the frontal and posterior scalp regions.

subset of these. First, as correlations are very common across neighbouring electrodes, we selected a set of 10 electrodes that span the recording coverage (Figure 2.3). Second, we averaged the signal for each electrode over 100 ms time bins, from 0–1200 ms post-stimulus onset. The resulting EEG signal consisted of 10 spatial $\times$ 12 temporal = 120 features.

When using multiple EEG features to make predictions, the classification rule is not known a-priori (unlike as above). However, we can learn this through predictive modelling. We used two models, linear discriminant analysis (LDA; Fisher, 1936) and linear support vector machine (SVM; Cortes & Vapnik, 1995). In general, linear models are advantageous because these are easy to interpret; the weight of a feature in the model indicates its relative importance in the classification. Each model has a set of parameters, values of which are set through examples, also known as "training set". Once trained, the model can generate examples on its own, thus it can be used for predictions for unseen examples, also known as "testing set".

It is crucial to test the model on unseen examples; for the model could be too specific to the training examples, often by capturing the noise in it (also known as "overfitting") and thus cannot generalize. To reduce overfitting, the weights of the features in the model can be scaled, also known as "regularization." We used a regularized LDA classifier (fitcdiscr, MATLAB 2018a) where the covariance matrix was calculated as: $\hat{\sum}_\gamma = (1-\gamma)\hat{\sum} + \gamma \text{diag}\hat{\sum}$, where $\hat{\sum}$ is the empirical, pooled covariance matrix for the two classes and γ is a regularization parameter, lying between 0 to 1. SVM uses support vectors to draw hyperplanes that discriminate between the two classes. The support vectors are examples (here, EEG trials) that maximize the distance between the classes. We can reduce the chance of overfitting of an SVM model by setting a penalty (the box constraint parameter of fitcsvm; MATLAB R2018a) for mis-classifying examples that are on the class boundary. The default value for box constraint is 1 and in general smaller values allow for more regularization. In this study, for all LDA models, we set $\gamma = 1$, i.e., the maximum. For SVM, since a fixed maximum or minimum for the box constraint parameter does not exist, we chose a value that is reasonably smaller than the default value of 1, we set box constraint = 0.05. Importantly, the choice of regularization parameter values were independent of the test sets used to evaluate performance of the classifiers. Note that it is also possible to tune the regularization parameters for individual models. However, this did not alter the results substantially (see Figure 2.7).

We tested model performance through 10-fold cross-validation. Trials were randomly

split into ten equal-sized folds, with nine folds being used to "train" the model and the remaining fold to "test" it. This was repeated ten times, ensuring that each trial was once used as a test trial. Cross-validation folds were stratified, such that the number of examples for the two classes were the same across all training folds. For each trial in a test fold, the trained classifier computed a "score," which readily translates into the posterior probability of that trial belonging to each class. Probability estimates across all trials for a test fold were then sorted and set as thresholds for calculating the corresponding true-positive and false positive rates, to trace the ROCs and compute the AUC. Average AUC across the 10 test-folds was used as the final estimate of the classifier performance.

Note that classifier success can also be evaluated by "accuracy", calculated as: $\frac{(TP+TN)}{(P+N)}$; TP = true positives, TN = true negatives, P = positives and N = negatives. However, in our data, the two classes, hits and misses, were imbalanced (described in detail in the next paragraph). Since accuracy does not take false alarms into consideration, for imbalanced sets, it is possible to achieve very high accuracy when the classifier has a bias to predict the over-represented class. Thus, we did not use accuracy as the measure for classifier performance.

Class imbalance In our data, hits were more frequent than misses, making the training sets class-imbalanced. This could lead to the classifier getting biased towards learning more about and predicting more frequently the over-represented class. If this was true, rebalancing the classes either by undersampling the over-represented class or by oversampling the under-represented class can be helpful. Due to the small sample size of our data (≤ 225 in total per participant) we did not use undersampling. Instead, we used the Synthetic Minority Oversampling Technique (SMOTE; Chawla et al., 2002; Arora et al., 2018) to create new examples from the existing under-represented class examples. To create a new example, the algorithm 1) randomly selects an existing example from the under-represented class, 2) randomly selects one example from its k-nearest neighbours (from the same class), 3) calculates the distance between the two chosen examples, 4) adds a random number between 0 and 1 to this distance and 5) adds the distance (with added random noise) to the first chosen example. The new example, created this way, lies in between the original example and its chosen neighbour. We set the number of nearest neighbours in the SMOTE algorithm to 4 but note that the first nearest neighbour is the example itself. Thus, the effective number

of nearest neighbours considered for each example was 3. Synthetic minority samples were computed until the total number of examples in the two classes matched. Importantly, we only used SMOTE to balance the training sets. If SMOTE is used to balance the entire dataset, it is possible to end up with very similar trials in the training and testing sets, creating a double-dipping problem.

Cluster analysis of LDA weights For LDA, we can assess the importance of a feature from its coefficient or weight in the model. To check if any pattern existed in the distributions of feature-weights across participants, we performed a cluster analysis in MATLAB (R2018a) using the k-means algorithm (kmeans function from the Statistics and Machine Learning toolbox; Martinez, Martinez, & Solka, 2017). For a specified number of clusters, n, the algorithm minimizes the within class variance or the sum of distance of each point in a cluster from the centroid of the cluster. We ran the cluster analysis separately for 2, 3, 4 and 5 clusters. To avoid local minima, each clustering solution was minimized over 100 replications. For each clustering solution, we calculated the following distance measure for each participant (using the function silhouette in MATLAB R2018a):

$$S_i = \frac{(y_i - x_i)}{\max(x_i, y_i)}, \tag{2.1}$$

where x_i is the average of all distances from the i^{th} participant in one cluster to all other participants in the same cluster and y_i is the minimum of the average distances from the i^{th} participant to all other participants in all clusters other than its own cluster. This measure can range from -1 (indicating probable wrong assignment of a participant in a cluster) to 0 (participant can belong to either of the neighbouring clusters) and up to 1 (participant is distant from the neighbouring clusters). A set of 2 clusters was found to be the best possible solution, with the highest average value for this measure (0.11) across all participants [see Figure 2.10 for a visual representation of the distance measures across all participants, separately for the cases of 2, 3, 4 and 5 clusters]. To visualize the feature-weight pattern for each cluster, we used spline-interpolated topographic plots, created by the topoplot function of the EEGLAB toolbox (Delorme & Makeig, 2004). An inverse distance-weighting interpolation was used. This means that feature weight values for electrodes that were not used in the classification, were calculated from the weighted averages of the same for the electrodes used in the classification.

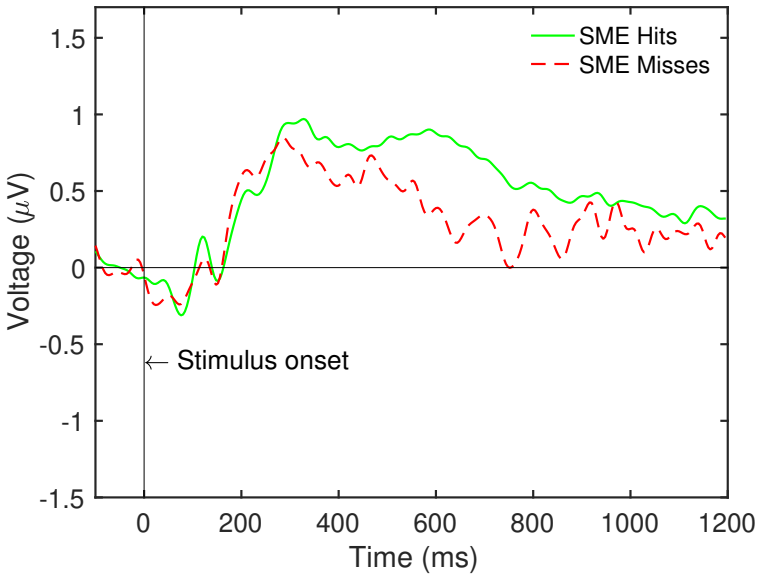

Figure 2.4: Grand averaged ERPs at electrode Pz for subsequently remembered (hits) and forgotten trials (misses).

All analyses were done using in-built and custom written functions and scripts in MATLAB R2018a. Specific functions from the Statistics and Machine Learning Toolbox (Martinez et al., 2017) were also used. Although the classification problem was set up for each participant individually, to gauge overall success of the methods, one sample t-tests (against chance level, 0.5) were done. We also carried out Bayesian t tests using a MATLAB function by SamPenDu (2015). The Bayes factor is the ratio of Bayesian probabilities for the alternative and the null hypotheses; $BF_{10} = \frac{p(H_1)}{p(H_0)}$. By convention (Kass & Raftery, 1995), $BF_{10} > 10$ provides strong evidence for the alternate and $BF_{10} < 0.1$ provides strong evidence for the null. For $BF_{10} > 3$ and $BF_{10} < 0.3$ there is some evidence for the alternate or the null, respectively. Effect sizes of the classifiers were estimated from the 95% confidence intervals. To ensure that our results can be reproduced over multiple runs of the scripts, a pseudo-random number generator algorithm was specified in MATLAB R2018a (Mersenne twister, seed = 0).

2.3 Results

We start with the traditional ERP analysis of the subsequent memory effect. Figure 2.4 presents these ERPs at electrode Pz, averaged across all participants ($N = 62$), whereby hits appeared to be more positive than misses. Paired t-tests between the mean voltage for

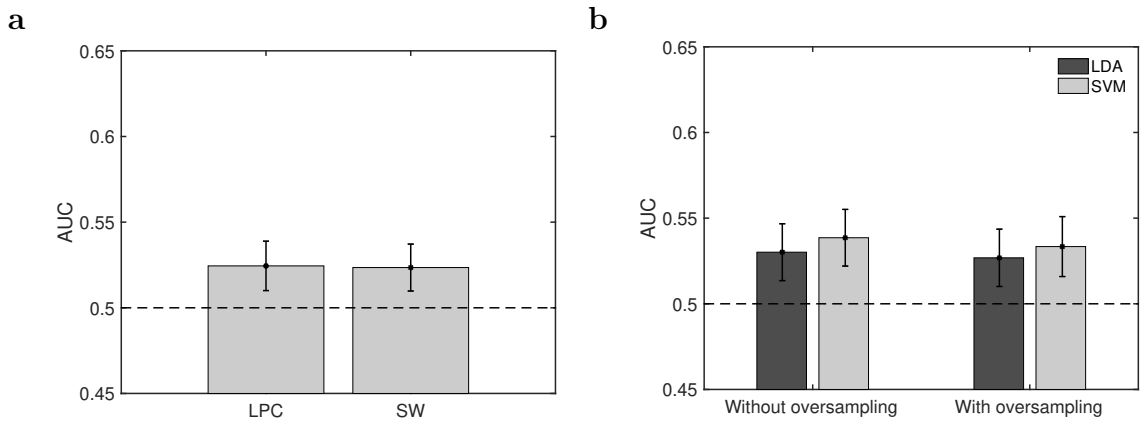

Figure 2.5: a. Classification based on SME ERPs: LPC and SW (computed from electrode Pz). Maximum AUC observed was 0.69 for both LPC and SW (for the same participant). b. Multivariate classification with LDA and SVM (left) and with oversampling to produce balanced classes (right). Maximum AUCs observed were 0.69 for LDA and 0.73 for SVM (same participant for LDA and SVM and also same as above). With balanced classes, maximum AUC for both LDA and SVM was 0.69 (same participant for LDA and SVM but different from above). Error bars are 95% confidence intervals. Dashed black line denotes chance level (0.5).

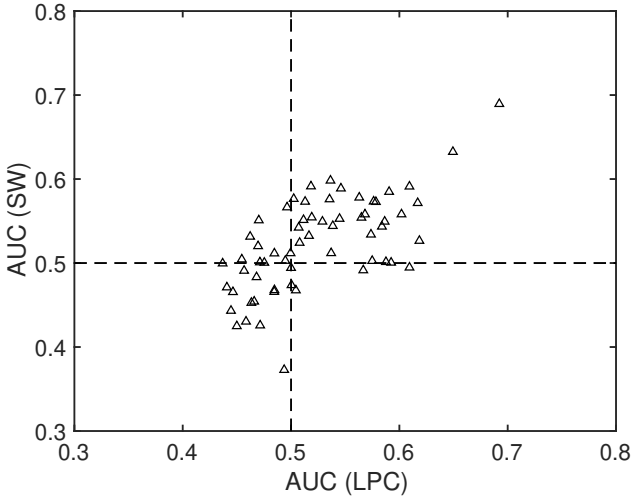

Figure 2.6: Correlation between AUCs for LPC and SW. Dashed lines denote chance.

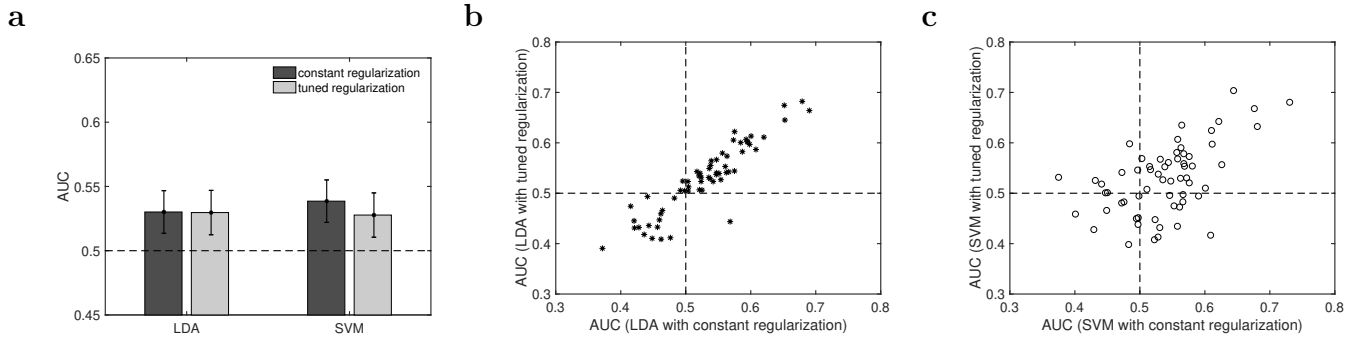

Figure 2.7: Effect of tuning the regularization parameters gamma of LDA and box constraint of SVM. We used a nested cross validation procedure. For the outer cross-validation, data was randomly partitioned into 10 stratified folds, 9 folds being used for training and 1 for validation. Then, the training data was subjected to an inner 9 fold stratified cross validation to tune the regularization parameter. For each training set of the inner cross validation, separate LDA models were trained for gamma = [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]. Similarly, for SVM, separate models were trained for box constraint = [0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100]. Then performance for these models were computed for the test folds of the inner cross validation. Value of the regularization parameter corresponding to the model with best performance was selected. Then this value was used in the model for the training data of the outer cross validation and then tested with the left out validation set. Finally, AUCs were averaged across the 10 validation sets. a. The overall effect of tuning the regularization parameters for each model. b. and c. AUCs for individual participants with constant and tuned regularization parameters.

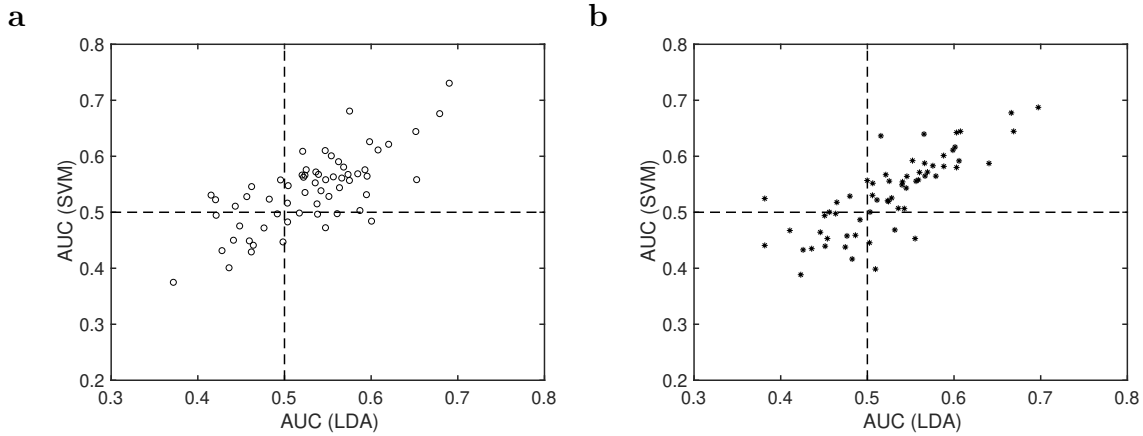

Figure 2.8: Correlation between AUCS for the two classifiers (LDA and SVM) with (a) and without (b) balanced classes for training. Dashed black lines denote chance.

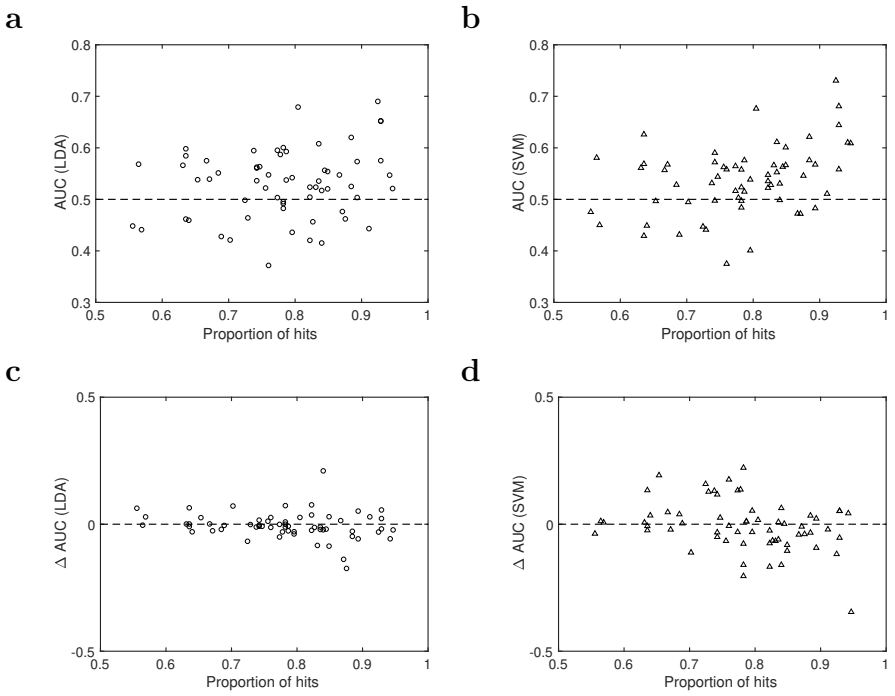

Figure 2.9: Relationship between classifier performance (AUC) and proportion of hits for LDA (a) and SVM (b). Percent change in classifier performance (ΔAUC) after oversampling, separately for LDA (c) and SVM (d).

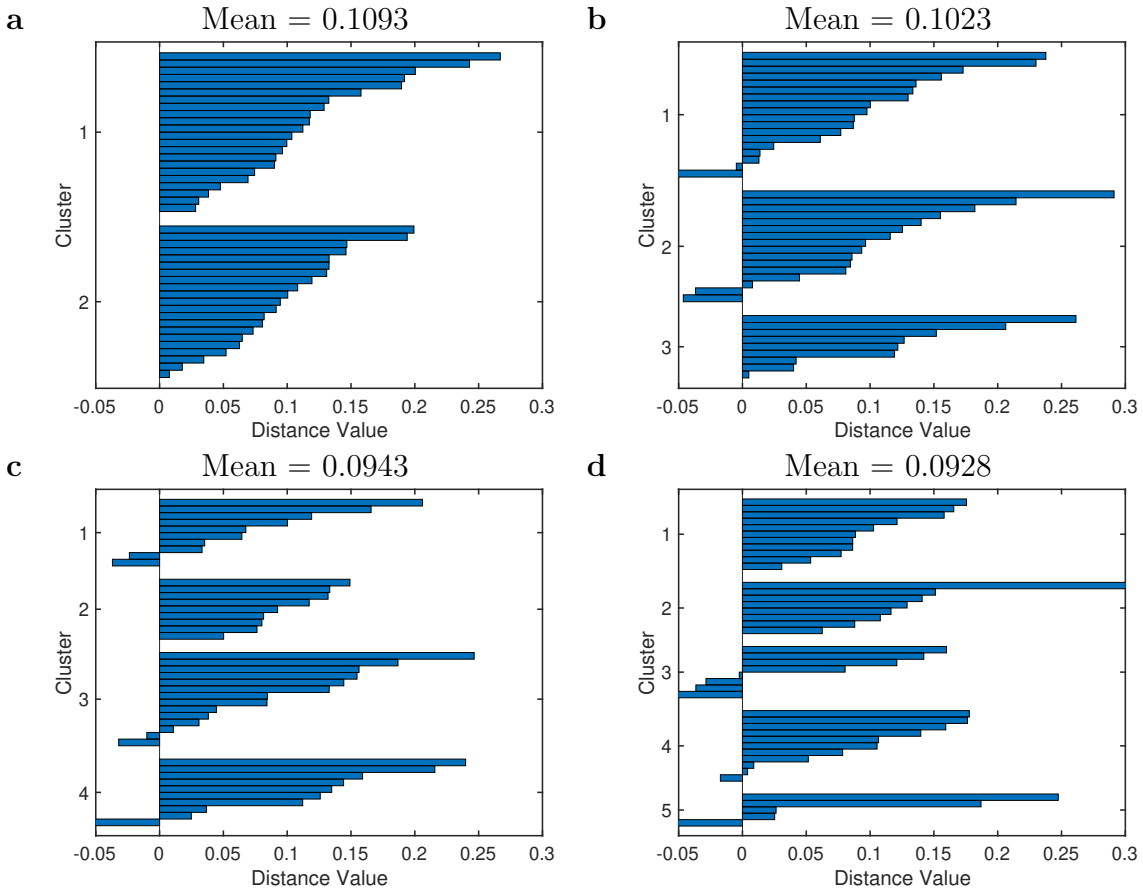

Figure 2.10: Determining the correct number of clusters for the cluster analysis of LDA feature weights. Each plot shows the distance measure for each participant for their respective clusters. Average distance scores across all participants are listed on top of the plot. For a set of two clusters (a), this measure was the highest. Also, all participants show positive distance scores for a set of two clusters.

hits and misses for the LPC was significant; $t(61) = 2.89,\ p < 0.05$. The difference was also significant for the early SW; $t(61) = 3.04,\ p < 0.005$. Note that these effects were comparable to those reported by previous studies. For example in Paller et al. (1987), the reported F ratio for the LPC was 8.6, thus the corresponding t-statistic can be estimated to be 2.93 (i.e., square-root of the F ratio), which is very similar to the present study. However, the ERP effect was not significant for the late SW; $t(61) = 1.82,\ p = 0.07$. We also calculated the Bayes factor, BF_{10}, which showed some evidence for the subsequent memory effect for the LPC ($BF_{10} = 6$) and the early SW ($BF_{10} = 9$) but was inconclusive for the late SW ($BF_{10} = 0.7$).

Next, we tested if these known SME ERP measures could predict subsequent memory for individual trials. Since the ERP effect for the late SW was not significant, we did not include it in this analysis. Accordingly, from here on, we refer to the early SW simply as SW. We found that for both ERP measures (Figure 2.5), AUCs (across all participants) were significantly above chance (0.5), $t(61) = 3.31,\ p < 0.005,\ BF_{10} = 18$ for LPC and $t(61) = 3.35,\ p < 0.005,\ BF_{10} = 19$ for SW. However, in each case, the 95% confidence intervals for the AUCs were close to chance; [0.51 0.54] for both LPC and SW. Also, across participants, AUCs were significantly correlated between the LPC and SW measures, $r(60) = 0.65,\ p < 0.0001$, (Figure 2.6). This could be due to the general temporal auto-correlation property of the EEG signal. In sum, classification of single trials from the study phase using a-priori measures achieved small but significant success.

Next, we tested if multivariate brain activity from the study phase, as measured with EEG, could predict subsequent memory and if it can do so better than the individual SME ERPs. As noted in the methods section, we selected a set of 10 electrodes and 12 time-samples, i.e., 120 features in total. We used two linear classifiers: LDA and linear SVM, along with a stratified 10-fold cross validation technique. AUCs were averaged across the 10 folds. To reduce chances of overfitting, we used regularization; the regularization parameters were set to be constant across the models (also, optimizing these parameters for individual models did not alter our results, see Figure 2.7). Across participants, the AUCs for both LDA and SVM (Figure 2.5b, left) were significantly better than chance, $t(61) = 3.54,\ p < 0.001,\ BF_{10} = 33.54$ for LDA and $t(61) = 4.55,\ p < 0.0001,\ BF_{10} > 500$ for SVM. The corresponding 95% confidence intervals were [0.51 0.55] for LDA and [0.52 0.56] for SVM. Also, pairwise one tailed t-tests showed that SVM performance was significantly greater than

the SME ERP based classifiers [SVM versus LPC: $t(61) = 1.83,\ p < 0.05, BF_{10} = 1.28$; SVM versus SW: $t(61) = 1.76,\ p < 0.05, BF_{10} = 1.13$]. However, this was not true for LDA [LDA versus LPC: $t(61) = 0.62,\ p = 0.27, BF_{10} = 0.24$; LDA versus early SW: $t(61) = 0.70,\ p = 0.24, BF_{10} = 0.26$]. Given that the multivariate models had more degrees of freedom than the SME ERP based classifiers, these results suggest that overall, the time domain EEG signal during the study phase is only marginally predictive of subsequent memory success. Moreover, predictive success was positively correlated between LDA and SVM (Figure 2.8a), $r(60) = 0.74,\ p < 0.0001$, suggesting that participants who were easier to classify by one method were also easier to classify by the other.

Notably, for both LDA and SVM, a small subset of participants were found to have AUCs far below chance (see Figure. 2.8a). This is possible, for the assumption of a symmetric null distribution for the classifier performance may not hold in the case of small sample size data with small effect size (Jamalabadi, Alizadeh, Schönauer, Leibold, & Gais, 2016). In that case, non-parametric tests may be better suited. Following up on this, for each participant we conducted a Mann Whitney U test between the AUC values for all of the 10 folds and chance (0.5). Then, we calculated the z transform of the U statistic. Finally, we used t-tests to check if the z scores across all participants were significantly different from zero. This showed that the z-scores for both LDA and SVM were significantly positive, [LDA: $t(61) = 2.55,\ p < 0.05, BF_{10} = 3$, SVM: $t(61) = 3.13,\ p < 0.005, BF_{10} = 11$]. This confirms that even if the assumption of symmetry for the null distribution is relaxed, the LDA and SVM classifiers in our study were overall better than chance.

Imbalanced classes might have challenged classifier training. Alternatively, participants with better memory may have a greater signal-to-noise ratio (SNR) that the classifier could identify. Across participants, a weak positive trend (Figure. 2.9a–b) was observed between AUCs and the proportion of hits. This trend was significant for SVM, $r(60) = 0.38, p < 0.005$ but not for LDA, $r(60) = 0.21, p = 0.09$. We also calculated the sensitivity index or d' of participants performance, which showed similar results as the proportion of hits [LDA: $r(60) = 0.14,\ p = 0.29$; SVM: $r(60) = 0.28,\ p < 0.05$]. To investigate if the imbalance between the trial numbers for hits and misses influenced our classifier results, we balanced the trials by oversampling the misses with Synthetic Minority Oversampling Technique or SMOTE (see Methods; Chawla et al., 2002). AUCs (Figure 2.5b, right) were yet again significantly above chance [LDA: $t(61) = 3.13,\ p < 0.005,\ BF_{10} = 10.91$, SVM: $t(61) =$

3.72, $p < 0.001$, $BF_{10} = 57$]. However, the 95% confidence intervals were not better than that without oversampling (LDA: [0.51 0.54]; SVM: [0.52 0.55]). Predictive success remained positively correlated across LDA and SVM, $r(60) = 0.80$, $p < 0.0001$ (Figure 2.8b). Thus, while imbalanced classes often pose a challenge to classifier training, in this case, it could not account for the relatively small prediction rate. Instead, participants with better recognition memory appear easier to classify (see Discussion for implications of this). The positive trend between classifier performance and proportion of hits was also observed after the classifiers were trained with balanced classes [LDA: $r(60) = 0.14$, $p = 0.29$; SVM: $r(60) = 0.13$, $p = 0.31$]. For SVM, when participants with very low AUCs (< 0.45) were excluded, this trend was significant, $r(51) = 0.27$, $p < 0.05$. However, classifier performance did not correlate with d' in this case.

For participants with LDA AUCs above 0.5 ($N = 43$), we wondered which features were deemed more important by the classifier for the classification. A cluster analysis of the LDA feature-weights revealed two subgroups of participants with distinct patterns (see Methods and Figure 2.10). $N = 22$ participants were found to be in cluster 1 and $N = 21$ in cluster 2. Figure 2.11 shows the topographic plots for the LDA feature-weights, averaged across all participants in each cluster and for three different time windows: 0–100 ms, 501–600 ms and 1001–1100 ms (see Figures 2.16 and 2.16 for the full version, i.e., for all the time-windows). For cluster 1, for the very early 0–100 ms time window, greater feature weights were observed on the left and right parietal regions. On the other hand for cluster 2, for the same time window, greater feature weights were observed in the fronto-central region. Given that these earlier time windows are more likely to reflect perceptual processing, one possibility is that these differences in feature weight patterns are indicative of the potential difference in attentional mechanisms between the two clusters. For a later time window, 501–600 ms, which is closer to the onset of LPC activity, only cluster 1 showed greater weights for the central parietal scalp region, whereas for cluster 2, greater weights were observed more widely in the right frontal and parietal regions. Thus, the topographic plot for cluster 2, for the 501–600 ms did not resemble the posterior positivity feature observed in the SME ERP analysis of this data set (see Chen et al., 2014). For an even later time window, 1001–1100 ms, the patterns for the two clusters were almost orthogonal; cluster 1 showed greater weights in the left parietal region whereas cluster 2 showed greater weights in the frontal and slightly left parietal region. It is possible that this difference is indicative of potentially

different spontaneous study strategies between the participants of the two clusters.

We were also curious as to whether the standard SME ERP effects might be different for the two clusters. Investigating this, follow up analysis of the corresponding ERPs at electrode Pz (Figure 2.12) showed a general trend for hits to be more positive than misses (i.e., the classic subsequent memory effect) for both clusters. However, this trend was clearly more pronounced for cluster 1 than cluster 2. We conducted a 2×2 ANOVA on mean LPC amplitude with the within subject factor memory success (hit versus miss) and between subject factor cluster (1 and 2). This revealed a significant interaction between the two factors, $F(1, 41) = 15.72,\ p < 0.001,\ \eta_p^2 = 0.28$, whereas the LPC effect was significant for cluster 1, it was not so for cluster 2. The same ANOVA design on the mean amplitude for SW, also showed similar results.

The average LDA AUC for cluster 1 was similar to that of cluster 2 [mean ± SD of AUC for cluster $1 = 0.56 \pm 0.05$; cluster $2 = 0.57 \pm 0.04$] and the average proportion of hits was comparable between clusters 1 and 2 [mean ± SD for proportion of hits for cluster $1 = 0.79 \pm 0.11$; cluster $2 = 0.80 \pm 0.08$]. The d' values were also comparable between the two clusters [mean ± SD for d' for cluster $1 = 2.13 \pm 0.62$; cluster $2 = 2.09 \pm 0.87$]. Overall, this could suggest that there may be *at least* two different types of feature patterns that could form the basis for predicting memory.

2.4 Discussion

The subsequent memory approach is often referred to as identifying brain activity "predictive" of memory. However, limited attempts have been made to test this claim with actual predictive models. Here, using signal detection theory, we showed that two, previously identified, SME ERPs, namely, the LPC and the SW, could indeed predict memory (hit or miss) for individual trials in a word recognition task. However, across participants ($N = 62$), the success rate was small. Considering the SME approach is limited by many factors, such as planned comparisons and trial averaging, the small success may be expected. Also, multiple processes could be at play for memory judgments in a recognition task, each associated with different sources of neural activity. Thus, instead of single ERPs, analysis of patterns in the multivariate EEG waveform at study may fare better at predicting memory. To test this, we employed machine learning classifiers (LDA and linear SVM), which are well suited to

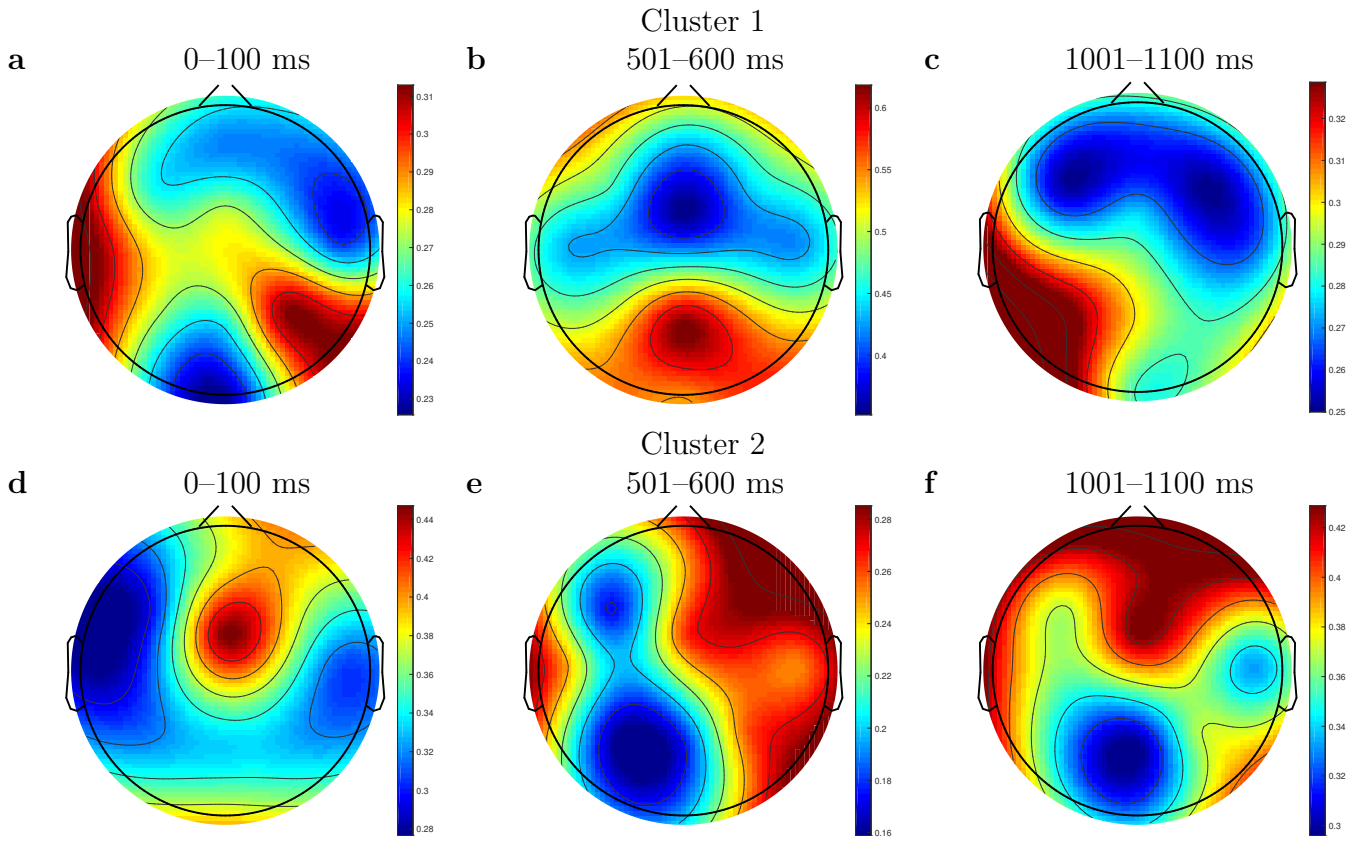

Figure 2.11: Cluster analysis of feature weights for all participants with LDA AUC > 0.5. A set of two clusters best explained our data ($N = 22$ for cluster 1 and $N = 21$ for cluster 2). (a–c) refers to cluster 1, (d–f) refers to cluster 2. Colors are range scaled. Note that the color scale varies across panels. See Figures 2.16 and 2.17 for full version of this figure.

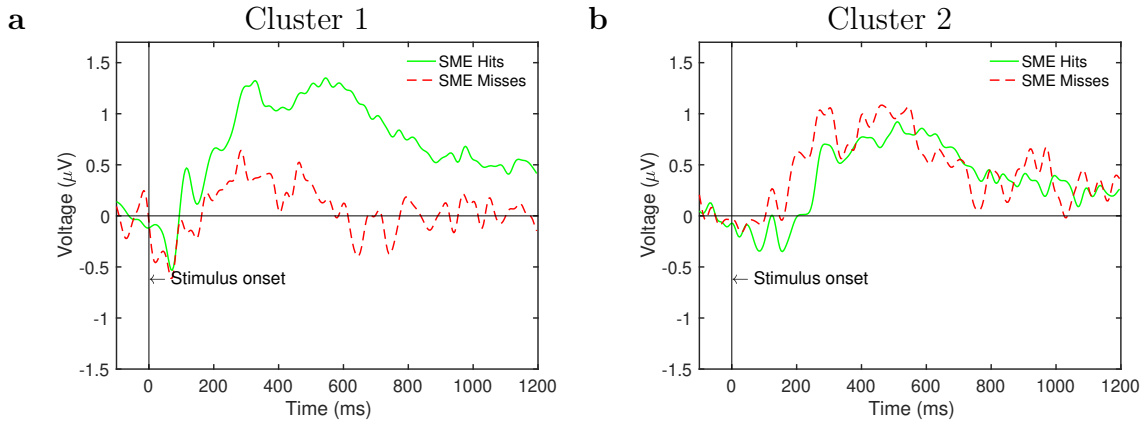

Figure 2.12: ERPs at Pz for the two clusters obtained through k-means clustering of LDA feature-weights.

analyze multivariate structures. These models were used to learn memory relevant patterns from a set of 120 spatio-temporal EEG features from individual study trials. Both LDA and SVM achieved significant success in predicting memory, albeit still with a small success rate. Since generalization was also accounted for in LDA and SVM with ten fold cross validation, the success of these models further strengthens the possibility of predicting subsequent memory from EEG activity at study. However, when comparing LDA and SVM performance with that of the LPC or SW based classification, only SVM showed a small significant improvement. Thus, despite the considerably greater degrees of freedom, these models did not offer an obvious improvement over classification with LPC or SW alone. But, interestingly, exploratory analysis on the features of importance to the LDA classifier showed that there were two subgroups of participants with seemingly different activity patterns. On average, one of these subgroups (cluster 1, see Figure 2.16) showed greater feature weights for the posterior scalp region, which is similar to the findings from the univariate ERP analysis of the same dataset (see Chen et al., 2014). It also agrees with previous SME ERP studies that have shown that memory success can be associated with a greater positive going signal over the parietal region (for a review, see Paller & Wagner, 2002),. However, for the other subgroup (cluster 2, see Figure 2.17), greater weights were observed in the frontal region. Further, post-hoc analysis of SME ERPs at electrode Pz, separately for the two subgroups, showed significant LPC as well as SW effects for cluster 1 but not for cluster 2. Interestingly, previous literature also suggests that a frontal slow wave may be invoked by associative processes whereas the posterior slow wave may reflect elaborate item-oriented processing (for e.g., see Kamp, Bader, & Mecklinger, 2017). Although we can not know this for sure, one possible reason for the involvement of the frontal region in cluster 2 could be that it reflects some associative strategies for learning, spontaneously undertaken by the participants in this group. Notably, it would not be possible to identify these subgroups without the classifier models. Thus, both the univariate and multivariate predictive analysis reported in the current study have their own merit. Below we discuss potential improvements and limitations towards predicting memory.

We sought out to understand the general level of challenge in predicting memory from EEG activity during the study phase. Unlike other approaches, where failed or less successful analyses might not be disclosed, so the degree to which best-cases are reported becomes impossible to judge, we report a systematic sequence of classification analyses, to avoid

apparently inflated success rates. We did not exclude participants based on their performance in the task, which is commonly done (Noh et al., 2014; Sun et al., 2016; Watanabe et al., 2011). Thus, although the 95% confidence intervals of our classifier success are modest for the aggregate, the regression (Figure 2.9b) suggests meaningfully large classification success rates. Also, to avoid any possible circularity in the analysis, we did not select the multivariate features based on the univariate SME results (Coutanche, 2013; Noh et al., 2014). Instead, we selected these features based on the general EEG knowledge (scalp coverage, correlations etc.), which substantially minimizes the chance of over-estimating the effect. Thus, our results provide a benchmark for the effect size for this type of classification to be compared against. This could improve with more fine-grained analysis, for example through other feature selection or feature reduction methods or even with the help of non-linear classifiers including state-of-the-art neural networks. Including EEG spectrogram features, which are more resilient to trial-to-trial latency fluctuations, may also lead to better performance (Ezzyat et al., 2017; Weidemann et al., 2019).

Notably, class-imbalance (hits versus misses) was common in our data set and this could have biased the training of LDA and SVM towards the over-represented class (hits). However, re-balancing classes offered no improvement, allaying such concerns, at least in our case. Alternatively, it is possible that participants with more hits also have high-SNR brain activity, which could have helped the classifier. We found some support for the latter, as SVM performance increased significantly as the proportion of hits increased (for similar evidence, see Arora et al., 2018). Whenever memory was close to chance, the corresponding brain activity may have had less information for the classifier to pick up on. Conversely, participants who performed better might, to some degree, have been those who (and whose brains) were more engaged in the task, producing higher task-relevance of their brain-activity.

Although we cannot know for sure why better-performing participants may be easier to classify, two causes come to mind. First, some lower-performing participants might have low motivation, a plausible possibility, given that participants did not self-select as research participants, but were recruited from via a course-based research participation pool, in exchange for partial course credit. There was no disincentive to speed through the experiment or disengage from the task. For such participants, brain activity may simply not be task-relevant. Second, participants who struggle with the task more, genuinely finding the task challenging, may have task-related brain activity that is more variable, or obscured by cognitive processes

related to frustration, or strategic exploration, etc. Both causes could lead to lower SNR. In future studies, this could be addressed by pre-calibrating the task for each participant to equate difficulty across the sample, and increase the level of motivation across participants, for example, through rewards. Both these modifications might produce substantially higher levels of classifier success as well across the sample.

One may conclude that due to individual variability in performance and likely in brain activity too, a large sample size, as in our study, is essential to obtain overall significant results with the classifiers. To test this idea, we estimated the minimum sample size we might have needed for the classifier analysis to succeed. With bootstrap techniques, we tested significant effects for the SVM classifiers for different sample sizes, ranging from 6 to 62 participants, which were selected at random and without replacement. For each sample size, we generated 100 sets of participants and for each set we calculated one sample t-test to check if the corresponding SVM AUCs were significantly better than chance. Then, across the 100 sets, we calculated the average effect or the probability of obtaining AUCs that were not overall significantly better than chance. This showed that the probability to obtain a non-significant effect for SVM decreases very sharply with increasing sample size (Figure 2.13), up to about 30 participants. For sample sizes greater than 30, this probability is very close to zero. Thus, we were not after a result that is only made possible by a large sample size. In fact, in many cases, a sample size of about 15 participants may be enough to obtain significant results (probability for a non-significant effect < 0.5, see Figure 2.13), provided the number of trials per participant is high or at least comparable to our study. Thus, it is conceivable that Sun et al. (2016) failed to find overall success with simple linear classifiers due to small sample size ($N = 9$). Interestingly, while some of the early, influential, SME ERP studies do not pass this sample size ($N > 15$) criterion (Brewer et al., 1998; Karis et al., 1984; Neville et al., 1986; Sanquist et al., 1980; Wagner et al., 1998), others do so (Smith, 1993; Otten & Rugg, 2001; Paller et al., 1987; Van Petten & Senkfor, 1996; Friedman, 1990). However, many of these also have considerably lower trial counts.

Many decades of behavioural research (Kahana, 2012; Humphreys, Bain, & Pike, 1989; Neath, 1998; Lewis, 1979) points to numerous factors that determine memory success, that should not be visible through the lens of study-related activity alone (for an alternate account, see Weidemann & Kahana, 2019b). Examples include competition from other items at retrieval, nature of the retrieval task (such as recognition, free recall, serial recall, cued

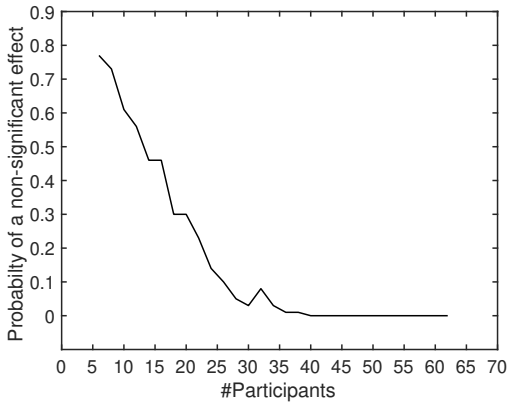

Figure 2.13: Effect of sample size on the overall significant results for SVM. With one sample t-tests, we calculated if SVM performance was significantly better than chance, for different sample sizes, ranging from 6 to 62 participants. For each sample size, participants were selected at random and without replacement. Further, for each sample size, we collected 100 sets of participants. Y axis shows the probability of obtaining a non-significant effect, calculated across these 100 sets and for each sample size.

recall, word-stem completion, word-fragment completion, lexical decision), retrieval time, output encoding, rehearsal and response criterion (recognition tasks). The serial positions of the items in the studied list can also influence subsequent memory (for example, primacy and recency effects) and are possibly reflected in brain activity as well (Talmi, Grady, Goshen-Gottstein, & Moscovitch, 2005; Rushby, Barry, & Johnstone, 2002; Sederberg et al., 2006). However, these factors are usually not accounted for in the SME approach. In addition, the Encoding Specificity principle (Tulving & Thomson, 1973) suggests that remembering will be more successful when there is a good match of context between study and test than when they mismatch (for an alternate account, see Nairne, 2002). Context could be spatial/environmental, temporal or internal mental or physical state (Howard & Kahana, 2002). This study–test contextual match is also overlooked with the SME. Brain activity indexed by the SME may also relate to experimental manipulations such as attention (Paller et al., 1987; Otten & Rugg, 2001; Summerfield & Mangels, 2006), intentional learning (Paller, 1990; Karis, Bashore, Fabiani, & Donchin, 1982), use of different learning strategies (Karis et al., 1984; Rugg & Curran, 2007), etc. Semantic congruity of the the to-be-remembered stimuli (Neville et al., 1986) as well as the type of the stimulus (for example, verbal, pictorial, abstract patterns etc., see Fabiani et al., 1990; Paller et al., 1987; Friedman, 1990; Van Petten & Senkfor, 1996) can also influence the SME. Also, study and test phases are temporally

distinct, but some aspect of brain activity may co-vary across these two phases (Chen et al., 2014), and test activity (Rugg & Curran, 2007) is also an important determinant of memory. Brain activity at retrieval may even be more reflective of important determinants of memory success (Weidemann et al., 2019; Polyn, Natu, Cohen, & Norman, 2005), including item-distinctiveness (LaRocque et al., 2013). Clearly, memory encoding is multifaceted and thus, a more extensive model that incorporates different cognitive measures as well as measures of brain activity may be more effective in predicting memory (Halpern et al., 2018).

Accordingly, it is likely that our current classifiers are "under-performing" their potential. One important factor missing from the SME approach and possibly influencing our classifiers is that retrieval performance is, to a large extent, competitive. Thus, probability of remembering an item not only depends on the corresponding EEG activity for that item at study but also on the EEG activity during the study of other items. Also, over the course of an EEG recording session, there is usually drift in the signal, mainly due to the electrodes drying out or sliding. Additionally, it is also possible that as the task progresses, the participant shifts their strategy or approach towards the task. All of these factors could influence the classification. To test this, with the LPC and SW classifiers, we calculated classification performance separately for each of the nine lists studied by each participant (Figure 2.14). Lists with all hits or all misses were not included. Indeed, the classification improved as the task progressed. Linear regressions between average AUC (across participants) for each list and list number (1 to 9) were significant, for both LPC [$F(1,7) = 6.34,\ p < 0.05$] and SW [$F(1,7) = 7.35,\ p < 0.05$]. Similar trends may also be possible for the multivariate classifiers, but due to the very small number of trials available per list for training the models, we did not follow up on that. However, performance measures such as d' or the proportion of hits for individual lists did not vary significantly with list number.

Also, given the wealth of research on distinctions between recollection- and familiarity-based retrieval, one important future direction could be to incorporate those distinctions into the classification— as indeed, has been done by some previous studies (Noh et al., 2014; Fukuda & Woodman, 2015; Noh et al., 2018; Liao et al., 2018; Sun et al., 2016). As with incorporation of other relevant variables (previous two paragraphs), this could improve classification accuracy. Two classifiers could be trained, one to classify based on a familiarity-like signal and one based on a recollection-like signal. The two classifiers could then be combined to produce a higher overall classification success rate. However, the subjective responses

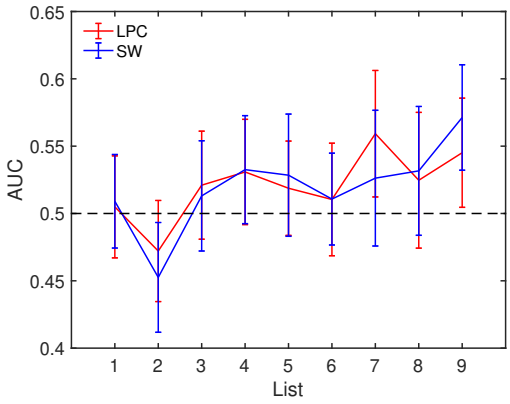

Figure 2.14: Classification of hit versus miss trials for each list in the task, based on the LPC and SW ERP measures. Error bars are 95% confidence intervals. Dashed line refers to chance performance. Lists with all hits or all misses were excluded.

distinguishing recollection versus familiarity might be variable, in themselves, and thus introduce noise into the classification. Moreover, Dunn (2008) showed that remember/know judgements, themselves, appear to be based upon a summation of recollection and familiarity evidence. Thus, alternatively, it could be more effective to let a multivariate classifier "discover" the two (or more) neural processes and their optimal summation weights.

Importantly, in the traditional ERP or other similar univariate analysis, brain activity is averaged over many trials to increase the signal to noise ratio (SNR). Then, measures from this averaged brain activity signal are computed for behavioural conditions of interest and are compared across participants. With this approach, we may be able to identify some components of brain activity relevant to that behaviour, it should also be considered that the brain itself does not compute such averages to produce the behaviour. Instead, this is produced by the firing of networks of neurons. In that sense, the classifiers, which learn from the multivariate pattern of brain activity specific to individual events, may be closer to the way the brain works than the traditional approach. However, since the classifiers are driven by the data only, it is also possible for them to learn relations that are different from what actually produces behaviour. Thus, obtaining a function-to-structure map of memory may also be inaccessible with present methods of obtaining brain-activity measures (Henson, 2005). The two different clusters of participants identified here could reflect individual variability in studying the same information, for example, use of different strategies. Classifiers could also be showing differences due to the word stimuli (frequency, imageability etc.).

Also, if the EEG activity from study can predict memory, then, hypothetically, it could

also be used to guide restudy, as attempted by Fukuda and Woodman (2015). But in their case, the restudy re-randomized the relationship between initial study-related EEG activity and eventual memory outcome. It is possible that when enforcing better learning for stimuli tagged as "likely-to-be-forgotten" by the classifier, stimuli that were initially more likely to be remembered become weakened. If the goal is only to be able to predict memory, it may be possible to find differences that lead to some classifier success, which is appreciated. But, to be employed in a memory training framework, we may need to isolate EEG activity that taps into truly effective encoding processes.

In doing so it may also be possible to find memory relevant neural activity patterns that not only generalize within the trials for one participant but also across multiple participants. This is an interesting future direction and very few studies have attempted at "between subject" prediction of memory (Liao et al., 2018; Koch, Paulus, & Coutanche, 2020). Specific to EEG, Liao et al. (2018) found some success with the test phase activity. However, their experiment requested additional subjective judgements from the participants, such as remember–know as well as source and confidence judgements. Accordingly, the between subject classifiers were set up to predict memory outcomes that were constrained to these additional judgments rather than simple old/new responses. Thus, although their results can not be directly compared to the current study, we were curious if between subject prediction for hits versus misses is possible with the study phase EEG activity in our data set. However, given the small success rates of the within subject classifiers, we suspected that this may fail. We tested this with a leave-one-subject-out cross validation procedure. Data from one participant were selected at random and used as the test set. Data from all other participants were used to train the classifier. We chose linear SVMs, as our results show these may be better than the LDA. The leave-one-subject-out cross validation was repeated 62 times, i.e., until data from each participant were used as the test set. This produced AUCs greater than 0.5 for 34 out of 62 participants, i.e., for about 54% of the total sample, thus it was not significant across participants, $t(61) = 0.98,\ p = 0.33, BF_{10} = 0.22$.

Also, with fMRI, Koch et al. (2020) were able to predict the average encoding pattern, between their participants. Following this idea, we checked if we could predict the average EEG waveform at study for hit and miss events for a participant, based on the same for the other participants. Once again, we used leave-one-subject-out cross validation and linear SVM classifiers. Also, since in this case each test set has only two trials (averaged waveform

for hit and miss), instead of calculating AUCs for individual participants, we pooled together the classifier scores across subjects to calculate the ROC and the AUC of the ROC. Further, we used 1000 bootstrap samples to calculate the 95% CI of the AUC. This produced an AUC of 0.64, along with a 95% CI of [0.54 0.74] (Figure 2.15). Thus, it was possible to predict the average waveform for hit and miss events between participants. However, this may not be too surprising as the variance of the miss waveform may, in general, be higher than the hit waveform, due to the disparity in their trial counts, as we have discussed before. The classifier may be able to learn based on this difference in variance. Regardless, this initial set of analyses suggest that there may be multiple interesting directions to follow up on in the future with between subject classifications.

In sum, SME ERPs such as the LPC and SW may not only be related to memory success at the aggregate level, but could also predict memory for individual trials, albeit with small effect size. Some increase in effect size was achieved by using more features of the study-trial activity, and through multivariate pattern classification. This also showed that two distinct patterns of activity could be related to subsequent memory success (see Figures 2.16 & 2.17). Methodological improvements to the classification analysis may be able to increase the performance even further (for example, by using more complex algorithms and/or spectrogram information) and will be addressed by future research. Also, it is possible that unlike the EEG signal, the SME measured by the fMRI may contain a better SNR to predict memory success for individual trials. Alternatively, it is also quite possible for the classification success to never approach the maximum possible outcome, due to the numerous cognitive factors that are known to significantly influence memory success, but are not directly taken into account in the subsequent memory approach. In that case, even a low, but above-chance, classification is important, and a small level of success is, in fact, expected.

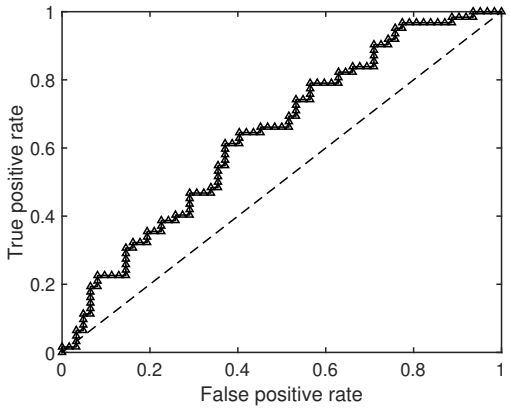

Figure 2.15: ROC curve obtained from between subject classification of the average EEG waveform at study for hit versus miss events, with linear SVM. Dashed line denotes chance.

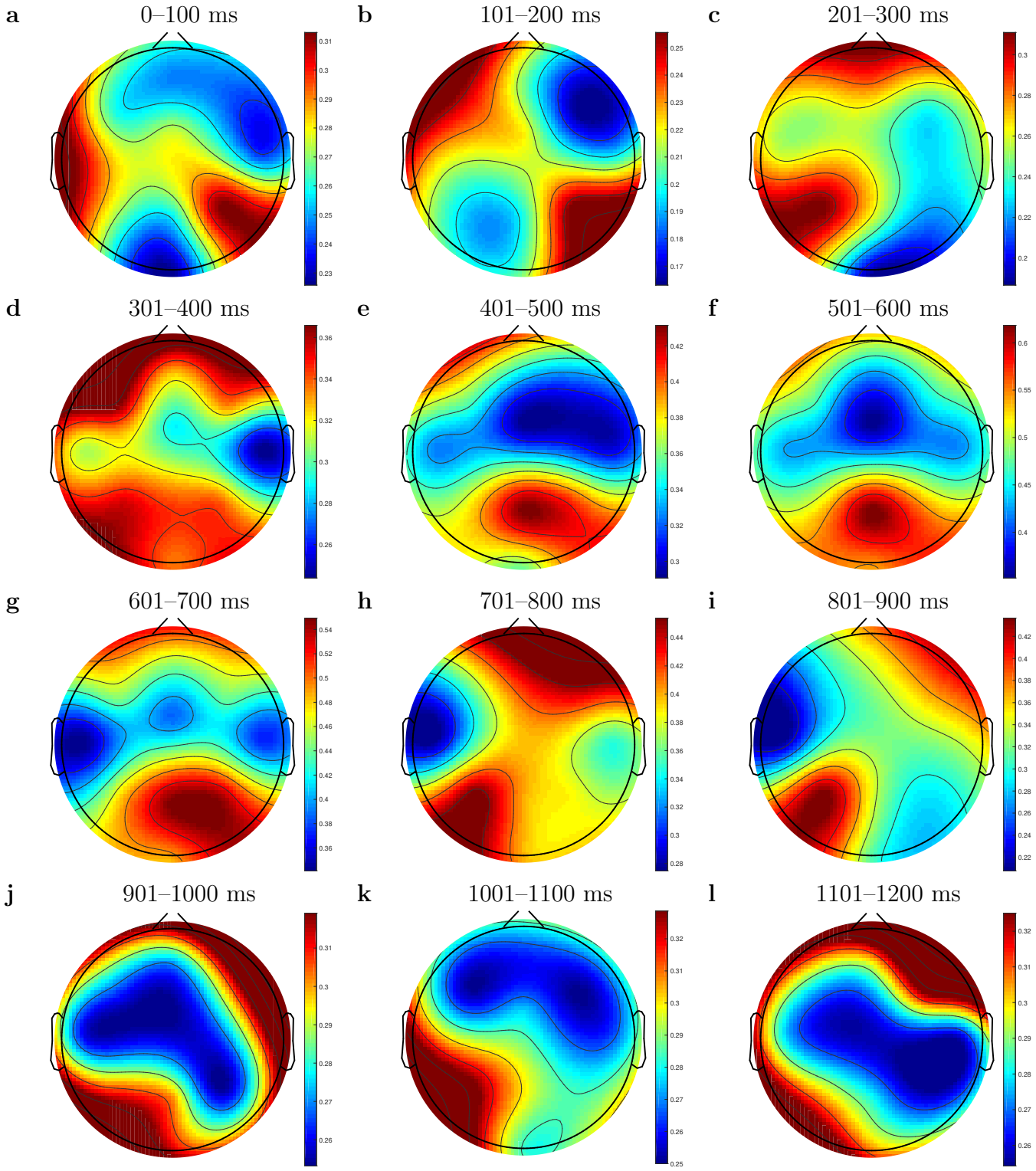

Figure 2.16: Topographic plots showing LDA feature weights averaged across all participants in cluster 1 ($N = 22$). Colors are range scaled and the scale varies across panels.

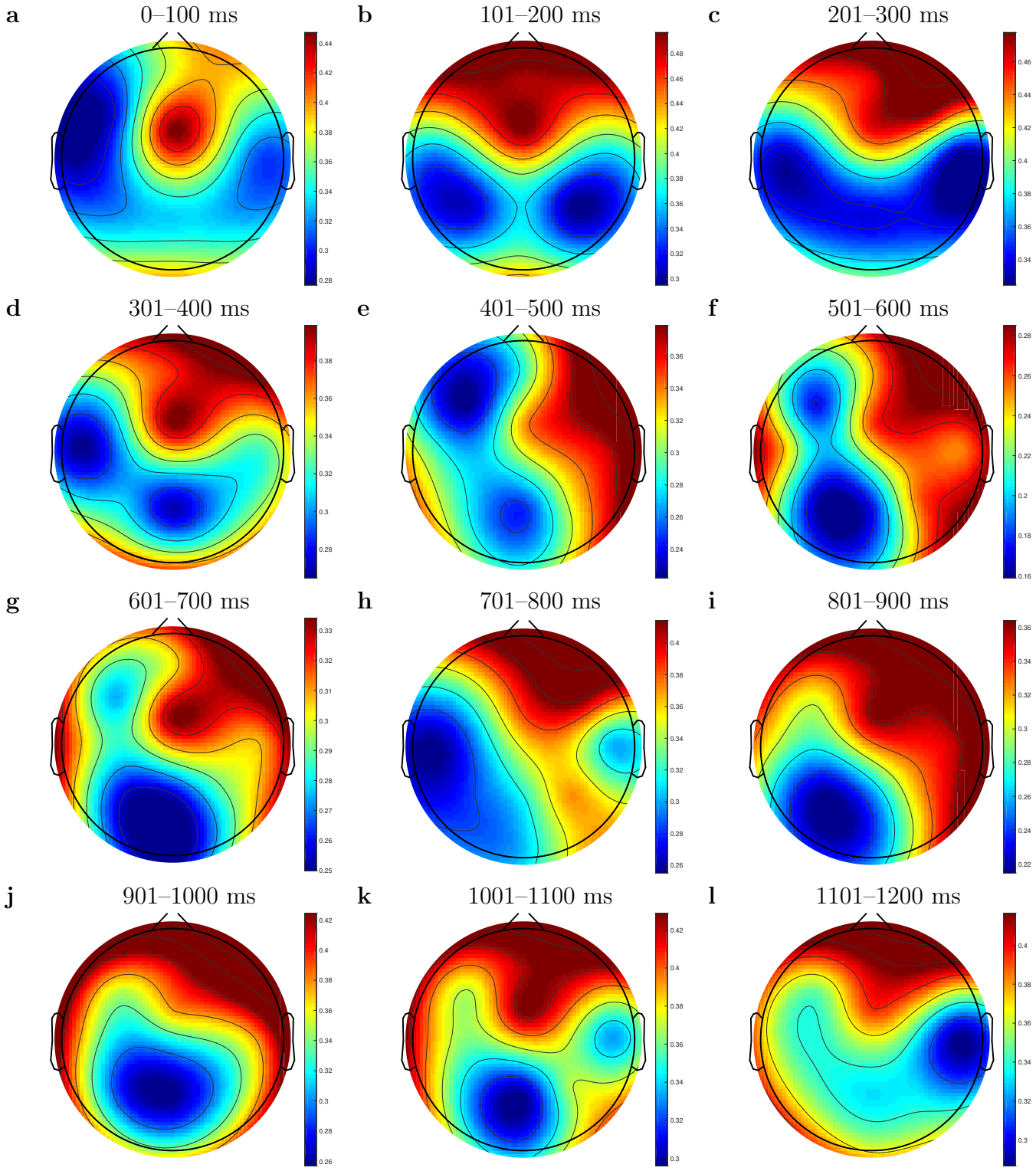

Figure 2.17: Topographic plots showing LDA feature weights averaged across all participants in cluster 2 ($N = 21$). Colors are range scaled and the scale varies across panels.

Chapter 3

Predicting memory from brain activity during the test phase of item recognition

Abstract

Memory judgments can be explained by cognitive processes that precede them in time and thus, by the brain activity that captures those processes. However, brain-activity signals are traditionally analyzed with planned-comparisons and descriptive methods, which can overestimate them, and also overlook subtle multivariate patterns. Predictive analyses could be a good complement to descriptive approaches and to find increasingly behaviourally relevant brain activity. We found that two previously-identified event-related potential measures of electroencephalographic recordings, FN400 and late parietal positivity (LPP)[1], present during tests of recognition, achieved modest success in predicting the memory outcomes, but their predictions did not correlate with each other, suggesting that the two signals may not contribute to common memory-variability. Further supporting this idea, multivariate pattern analysis of brain activity, which could identify combinations of different features, such as, the FN400 and LPP, achieved modest but significantly better success than FN400 or LPP voltage alone. Multivariate pattern analysis also showed that when decisions were reached fast, only one source of evidence might drive the judgments, in line with a single-process account; however, when the decisions took longer to reach, a dual-process account may be more accurate. Further, multivariate pattern analysis of brain-activity during the test-phase predicted memory better than that for study. However, when both study- and test measures were combined, predictions were similar to test only. Thus, building on previous univariate results (Chen et al., 2014), we found that test-related activity, which may be the result of predictive study-related activity, can predict memory outcomes more directly. Finally, performance of the multivariate classifiers, for both study- and test phases, correlated positively with participants' performance, suggesting that brain activity for better-performing participants may be more task-relevant. Overall, these investigations show that predictive

[1]Not to be confused with the emotion-related late positive potential (LPP; e.g., see Moran, Jendrusina, & Moser, 2013).

approaches are useful to evaluate behaviourally-relevant brain activity, and also to gain additional insights about the underlying cognitive processes.

3.1 Introduction

Cognitive processes that precede memory judgments in time can explain variability in memory judgments. Accordingly, analysis of brain activity that captures those cognitive processes could identify memory-relevant signals. For item recognition, previous investigations with electroencephalographic (EEG) recordings have identified two highly-replicated event-related potential (ERP) signals, elicited during the processing of the test items: the FN400 and the left-parietal positivity (LPP) (Chen et al., 2014; Friedman, 1990; Neville et al., 1986; Rugg & Nagy, 1989; Rugg, 1995; Rugg & Curran, 2007; Warren, 1980; Wilding & Rugg, 1996). Both FN400 and LPP show more positive amplitude for hits than correct rejections, commonly known as the old/new effect; FN400 and LPP also show more positive amplitude for hits than misses, commonly known as the retrieval-success effect (coined by Dolcos et al., 2005). Some have also reported more positive FN400 for false alarms than correct rejections (Finnigan et al., 2002; Wolk et al., 2006), which is in line with the idea that false alarms are erroneous memory-strengths for the test items that resemble the targets (we explain the memory-strength concept for recognition judgments in detail later). Thus, FN400 and LPP could index cognitive processes that support the recognition judgments. However, both old/new effect and retrieval-success effect have been obtained from traditional ERP analysis, which is based on planned-comparisons and descriptive methods, and runs the risk of overestimating the brain-activity signals, and also does not look into multivariate patterns of activity, which may be more relevant for explaining behaviour. In contrast, here, we consider a relatively stronger predictive framework to evaluate the behavioural relevance of ERP signals for recognition-memory, and evaluate multivariate classifiers as a way to identify increasingly behaviourally-linked brain activity.

Predictions over planned comparisons Whether memory can be predicted from brain activity, is not only an important question to advance theory, but also paves the way for memory-based applications, for example, in situations where overt behavioural responses cannot be obtained. Also, the question of predictability is not trivial, and may not be answered by the traditional ERP effects. ERPs are obtained after averaging over many trials, and thus only refer to the difference in the means for two or more conditions. However, at the individual trial level, the distributions of ERP amplitudes could be largely overlapping

(for example, see Figure 1.5c on page 20 of the Introduction Chapter), and thus, can be orthogonal to the inference drawn from the difference in the means. Further, trial-averaging favours signals with low trial-to-trial variability in the latencies, and washes out signals with greater variability in trial-to-trial latencies, which could also be behaviourally relevant (Luck, 2014). Although we cannot solve the problem of trial-to-trial variability in latency, it is possible to overcome problems due to inter-individual differences in latency with predictive analysis conducted for individual participants. Thus, if changes in FN400 or LPP amplitudes are indeed monotonic functions of memory outcomes, as suggested by the ERP effects, then, we should be able to discriminate between different memory outcomes using subject-specific thresholds in ERP amplitudes. Here, we test this idea using a signal-detection theory approach (Green & Swets, 1966).

The planned-comparisons approach is also subject to overfitting by capturing the noise (Bishop, 2006) that is generally present in recorded brain activity (e.g., task-irrelevant cognitive and neural processes). As a result, ERP research often reports findings that risk not being able to generalize (replicate) in future experiments operating on similar questions. Also, analysis of amplitudes for specific ERP waveforms restrict the discovery of subtle multivariate patterns of brain activity, which may better explain behaviour (Norman et al., 2006; Polyn et al., 2005). Overfitting can also be a problem for analyzing behaviourally-relevant brain activity for a different reason; for example, for memory tasks, there could be idiosyncratic reasons, and cognitive or neural processes correlated with them, for why a particular word is remembered, that do not generalize to very many other words. For example, the word Rocky may be memorable to someone because it reminds them of the movie Rocky; accordingly, brain activity related to the movie may *describe* why they remembered Rocky in one sample dataset, but it will not help explain why they might remember Pillow in a different sample.

In contrast to the descriptive methods, with predictive analysis, specifically with machine learning classifiers, testing for generalizability is standard practice. The models use different sets of observations to 1) learn the relationships between independent and dependent variables, also known as *training*, and to 2) forecast outcomes for unseen observations, known as *testing*. Thus, overfit models are likely to predict poorly for test data, and thus, both the model and the training parameters (or features) can be re-evaluated. Such tests of predictions, although still not tests for causality, can be stronger tests for finding behaviourally-

relevant brain activity, than planned-comparisons or descriptive approaches. Accordingly, we follow-up on the signal-detection theory inspired predictions with the univariate ERP amplitudes proposed above, with data-driven multivariate pattern analysis of brain activity present during the test-phase.

For both tests of predictions, with the univariate ERP measures and the multivariate activity, we asked the following four questions. First, can targets and lures be discriminated from brain activity, without restricting to correct responses only? Second, can participants' old/new responses be predicted from brain activity, distinguishing trials that were perceived-as-old (hits and false alarms) from those perceived-as-new (misses and correct rejections)? Third, does brain activity predict memory success for the targets (hits versus misses)? Lastly, can we predict false alarms and correct rejections based on brain activity measures? Taken together, we investigated the relevance of FN400, LPP, and multivariate test-phase activity in recognition-memory judgments, when subjected to tests of predictions for individual trials.

ERP analysis of our data, which was previously reported in Chen et al. (2014), showed significant old/new and retrieval-success ERP effects for both FN400 and LPP. However, participants' performance (d') as well as average response times (for hits) correlated with FN400 amplitudes (for the retrieval-success effect) only. Thus, FN400 may have been more relevant to the memory outcomes in this task than LPP. Accordingly, for this study, predicting memory outcomes could succeed better with FN400 than LPP amplitude. Alternatively, despite no correlation with behavioural measures, LPP could still predict memory. Also, as we explain below, FN400 and LPP are thought to index different cognitive processes that may contribute to the recognition-judgments independently (Rugg & Curran, 2007). In that case, with the multivariate pattern analysis, which could discover linear combinations of FN400 and LPP features, we may find greater predictive success than FN400 or LPP amplitudes alone.

Single- and dual-process accounts of recognition memory Researchers have long debated on how to interpret behavioural and neuroimaging or electrophysiological findings of the recognition task. Notably, mathematical models of memory frequently assume that successfully remembering one item while failing to do so for another is due to weaker memory-strength produced by the latter than the former (e.g., Shiffrin & Steyvers, 1997). Signal-detection theory (Green & Swets, 1966) has helped in explaining recognition judgments in

terms of memory-strengths— if strengths vary from item to item, we can assume separate (normal) distributions for targets and lures. It is also assumed that the mean strength (and in some models the variance too) for the targets is greater than that for the lures. Thus, if the memory-strength produced by a test-item exceeds some threshold value lying in-between the target and lure strength-distributions, the item is recognized as old, otherwise as new. Likewise, false alarms happen when the memory-strength produced by a lure item exceeds the threshold, which could happen in multiple situations, e.g., when target and lure distributions largely overlap or when the threshold is placed inside the lure distribution.

A single-process account, which is based on the signal-detection theory, posits that recognition judgments are driven by a unitary, integrated strength-based evidence (e.g., Dunn, 2008; Wixted & Stretch, 2004). On the other hand, a dual-process account posits that recognition judgments are driven by two (or more) independent sources of evidence, one of which is based on a strength-based familiarity signal, while the other is based on a process of recollection (Yonelinas, 1997, 2002). Interestingly, traditional ERP effects of FN400 and LPP add some support to a dual- than a single-process account. The common view is that FN400 indexes familiarity (Curran, 1999; Rugg & Curran, 2007) or conceptual priming (Voss, Lucas, & Paller, 2012) while LPP indexes recollection (Rugg & Curran, 2007; Wilding & Rugg, 1996). For example, FN400 amplitude is modulated by the different confidence ratings for old/new responses (Woroch & Gonsalves, 2010). Since confidence ratings are thought to be familiarity-driven (Voss & Paller, 2009), the above result may suggest that the FN400 is sensitive to different familiarity-based ratings. Results from the remember/know paradigm (Tulving, 1985), which is a derivative of the basic old/new task, add more support to this idea. In the remember/know task, participants are asked if, for the old response, they can 1) recollect specific details about the item or 2) can remember the item without recollection of specific details (or they they can remember the item with a sense of familiarity only), and respond 'remember' or 'know', respectively. LPP amplitude is modulated by whether the response is a 'remember' or 'know' (Rugg & Yonelinas, 2003; Curran, 2004, 1999), suggesting that LPP is sensitive to recollection of information. LPP amplitude is also modulated by correct and incorrect judgments of the source information (e.g., color or location of the studied items) (Guo, Duan, Li, & Paller, 2006; Woroch & Gonsalves, 2010). Since source details are considered to be some of the evidence on which recollection responses could be made, the above result also supports the idea that LPP indexes recollection-based processes.

Although the above findings of the FN400 and LPP are relevant to single- and dual-process accounts of memory, with ERP effects, it is not straightforward to measure the amount of neural evidence for memory-relevant information. Also, some have argued that the LPP occurs later than many of the responses, which makes it implausible as neural evidence that can actually drive the responses (Ally, Simons, McKeever, Peers, & Budson, 2008; Cabeza, Ciaramelli, Olson, & Moscovitch, 2008; Voss & Paller, 2008; Wagner, Shannon, Kahn, & Buckner, 2005; Woroch & Gonsalves, 2010). In contrast, classifier problems can be constructed to trace neural evidence for memory-relevant information more objectively. In a novel creative approach, Weidemann and Kahana (2019a) tested support vector machine classifiers (SVM; Cortes & Vapnik, 1995) on smaller time-intervals relative to the onset of the test item, both with moving averages of the signal (independent time-bins) and with cumulatively summed averages of the signal (cumulative time-bins), and they increased the time-intervals until the response was made. Their logic can be understood as follows: if recognition judgments are independently driven by different sources of evidence, then classifier performance for a specific time-bin should depend on the nearby source of evidence only, whereas performance for the corresponding cumulative time-bin will depend on all relevant sources up to that time. Thus, classifier performance for the cumulative time-bins will be greater than that for the independent time-bins. Alternatively, if memory decisions are based on a unitary, integrated signal then performance of SVM for the independent- and the cumulative time-bins should be very similar. In other words, if evidence at time t is no different than evidence over the interval $0 - t$, then a classifier using the signal at time t should perform similarly to a classifier using the signal over the interval $0 - t$. The results showed that SVM performance was very similar for each of the independent- and cumulative time-bins. Thus, the results from Weidemann and Kahana (2019a) were what one expects from a single-process account, and incompatible with a dual-process account. However, Weidemann and Kahana (2019a) analyzed spectrographic features (power as a function of time and frequency). The wavelet transform used to compute these features averages over a large time window around the time point in question. This might make their approach less sensitive to dissociations between the cumulative and independent time-binned classifications. Here we conducted a similar investigation with the time-domain features of EEG, which we thought could either reinforce the conclusion made by Weidemann and Kahana (2019a) or alternatively, might have the time-resolution to produce a result supporting

dual-process theory.

Cognitive processes at study Memory success for the target items in recognition tasks, or hits and misses, are also viewed as functions of cognitive processes during the study phase. In the "subsequent memory effect" technique, researchers have examined brain activity during the study-phase that supports difference due to memory success versus failure at later test (Brewer et al., 1998; Chen et al., 2014; Fabiani et al., 1990; Friedman, 1990; Karis et al., 1984; A. S. Kim et al., 2009; Paller et al., 1987; Sanquist et al., 1980; Smith, 1993; Wagner et al., 1998). Two highly-replicated ERPs at study, the late positive component (LPC) and the slow wave (SW) show greater amplitudes for subsequent hits than misses (Chen et al., 2014; Fabiani et al., 1990; Friedman, 1990; Karis et al., 1984; A. S. Kim et al., 2009; Sanquist et al., 1980; Smith, 1993). Based on similar motivations as the current study, previously we had looked into the predictability of the amplitudes of LPC and SW for subsequent memory success, and for the same dataset (Chapter 2; Chakravarty et al., 2020); LPC and SW amplitudes achieved small but significant success in predicting memory, which further supported their relevance in memory. In the same study (Chapter 2; Chakravarty et al., 2020), pattern analysis of the multivariate spatio-temporal EEG signals present during the study-phase, also predicted subsequent memory, and with similar average success as LPC- or SW amplitudes alone. Thus, overall we found support for the idea that memory outcomes (for the targets) can be viewed as predictive functions of brain activity during the study-phase, but the size of these predictions was small, suggested relatively small contributions of study-phase brain activity to subsequent memory variability.

It is arguable whether recognition-memory outcomes are better explained by cognitive processes at study or at test. Study- and test-related processes may also not be strictly independent of each other. For example, ERP analysis for the current task (Chen et al., 2014) found significant correlations across participants between trial-averaged SW amplitudes at study and LPP amplitudes at test, as well as for LPC amplitudes at study and FN400 amplitudes at test, when comparing between hits and misses. Thus, cognitive processes at study and test, as indexed by these ERPs, may share common variance in memory. However, some suggestions could make the case for greater importance of cognitive processes at test (e.g., Lewis, 1979). For example, consider that recognition memory depends on the ability to distinguish targets from lures. According to the signal-detection theory, classification of

a test item as old or as new depends on the match of the stored memory to the probe. Alternatively, test items could also be classified on the basis of their 'mismatch' with the lures (Criss, Wheeler, & McClelland, 2013). To visualize this idea, consider if lures are strikingly dissimilar to targets (e.g., different categories of objects), the recognition task can be performed extremely well, even without remembering specific items from study. Thus, lures play an important role in recognition-memory outcomes. Accordingly, test-related activity, which includes activity for the lure items, may explain more variability in memory accuracy.

Based on the above ideas, we compared between contributions of brain-activity features at study and at test, for explaining the memory outcomes, based on their predictive success. Thus, we compared between predictions of study- and test-related ERPs as well as between multivariate brain activity at study and test. If test-related activity, as discussed above, is more directly relevant for recognition judgments, we should achieve greater predictive success with it. Further, as mentioned above, results from Chen et al. (2014) suggested that cognitive processes at study and test may be linked through common variance in memory. Merging the two ideas together, it is possible that study-related processes lead to some variability in memory, which is retrieved by test-related processes, along with additional variability due to test only. Accordingly, test-related activity may explain a greater amount of variance in memory than study-related activity.

To test this idea, we trained the multivariate classifiers with brain activity features from both study- and test phases for the target items, in order to predict their memory outcomes (hit or miss). If our hypothesis is true then this *study+test* classifier should predict memory similar to the test-activity based classifier; and both the *study+test* and test-activity based classifiers will perform better than the study-activity based classifier. Alternatively, if despite a shared variance, study-related processes also contribute to the variance in memory that is not shared by the test-related processes, then the *study+test* classifier should perform better than both study-activity and test-activity based classifiers.

In sum, we examined the general scope for employing predictive analysis to investigate brain activity, measured by time-domain EEG features, that underlies recognition judgments. We interpret the predictive success on a background laid by results from the traditional ERP analysis, existing theories, as well as other classifier approaches relevant to understanding recognition memory.

3.2 Methods

3.2.1 Participants and experimental procedure

Data were from a total of 64 participants as reported in Chen et al. (2014) and in Chapter 2 (Chakravarty et al., 2020). Data from three participants were excluded for having 15% or more of the trials rejected due to artifacts (see below). The experiment was an item recognition task (see Figure 2.1 in Chapter 2, on page 40). Participants first studied a list of 25 words, presented on the screen one at a time. Word onsets were jittered (300–500 ms) and each word stayed on the screen for 1500 ms. After studying a list, participants completed a short math distractor task consisting of simple addition or subtraction problems. After that, they saw a list of 50 words, also presented one at a time. For each word, they made an old or new response by pressing specific keys (press 1 for old, 2 for new). Each test list contained 25 old and 25 new words. Each participant went through 9 study and 9 test lists, resulting in 225 and 450 study and test trials respectively. New words or lures were not repeated within a test list nor between different test lists.[2] The procedures were approved by a University of Alberta ethical review board.

3.2.2 EEG recording and pre-processing

The EEG signal was recorded with high-density 256-channel Geodesic Sensor nets (Electrical Geodesics Inc., Eugene, OR), in an electrically shielded, sound-attenuated chamber. It was amplified at a gain of 1000. The sampling rate was 250 Hz, impedance was kept below 50 $k\Omega$, and the vertex electrode Cz was the reference. We used the EEGLAB toolbox (`http://sccn.ucsd.edu/eeglab`; Delorme & Makeig, 2004) to pre-process the signal, which involved bandpass filtering to 0.5–30 Hz, average re-referencing and using the independent component analysis (ICA) method to discard artifacts like eye blinks, channel noise and muscle noise. Each trial included signal from 100 ms pre-stimulus onset to 1200 ms post-stimulus onset time-intervals (but see the description for the Vincentized analysis below, for which the epochs were extracted differently). Baseline, for each trial, was calculated by averaging the signal over the 100 ms pre-stimulus interval and was subtracted from all values in each particular trial. To detect epochs containing artifacts, an absolute voltage threshold

[2]Words in the study list were presented in the same order in the test list, with lures being presented at random positions in the test list.

of 200 μV was used. We also excluded trials based on a threshold of 25 μV point-to-point difference. For $N = 3$ participants, these thresholds resulted in the rejection of 15% or more of the trials (20%, 17.56% and 44.89% respectively) and thus data from those three participants were excluded. For all participants included in this study ($N = 61$), on average 1% of trials were rejected (min = 0, max = 9.78%).

3.2.3 EEG Classification

First, for each participant, the test trials were labeled by their memory outcomes: hits, misses, false alarms and correct rejections. Next, we constructed four classification problems: 1. old words or targets versus new words or lures, 2. words perceived/responded to as old (hits + false alarms) versus those perceived/responded to as new (misses + correct rejections), 3. correctly responded old words (hits) versus incorrectly responded old words (misses) and 4. correctly responded new words (correct rejections) versus incorrectly responded new words (false alarms). For classifications involving activity during the study phase, study trials were labeled by their subsequent memory outcomes: hits or misses. Below, we describe the classification analysis with the individual ERP amplitudes and with the multivariate EEG signal at test.

3.2.4 Classification based on ERPs at test

Two ERP features were considered: the amplitudes of FN400 and the LPP. Given the scalp-distribution of voltage for the FN400 (frontal) and LPP (left-parietal)(e.g., see Chen et al., 2014), for each trial, we calculated FN400 amplitude by averaging the signal from Fz over the 300–500 ms time-window post stimulus onset, and we calculated the LPP amplitude for each trial from the left posterior electrode P3 by averaging over the 500–800 ms window, post stimulus onset. Also, based on the ERP effects for the trial-averaged data, for both FN400 and the LPP amplitudes, we used the following rules for each of the classification problems mentioned above:

1. old/new: old trials are of more positive voltage than new trials.

2. hit/miss: hits are of more positive voltage than misses.

3. perceived old/new: perceived old trials are of more positive voltage than perceived new trials.

4. correct rejection/false alarm: false alarms are of more positive voltage than correct rejections.

Next, for classifying individual trials, we followed a signal-detection theory (Green & Swets, 1966) approach as described in Chapter 2 (Chakravarty et al., 2020). Specifically, for each participant, first we sorted the individual trials by their ERP amplitudes. Then we set variable voltage thresholds, whereby any trial with an average voltage above this threshold would be considered as part of the positive class (e.g., old for the old/new paradigm). Next, for each threshold we calculated the true positive rate and the false positive rate, which were plotted against each other to obtain the receiver operating characteristic (ROC) curve. Area under the curve (AUC) of the ROC was calculated using the *perfcurve* function in MATLAB (2018a). The AUC determined classifier success. For arbitrary predictions, AUC = 0.5, for a perfect classifier, AUC = 1 (Green & Swets, 1966).

3.2.5 Classification based on the multivariate EEG signal at test

Classification methods followed here were consistent with those used in Chapter 2 (Chakravarty et al., 2020); for both studies, we followed a generic approach in selecting the brain-activity features, and to analyze the classifiers. Our goal was to estimate the general level of challenge associated with using classifiers to predict different memory outcomes, leaving room for more fine-grained methods for the future.

Since EEG recordings contain thousands of features relative to much smaller number of trials, for computational simplicity and also to prevent the classifiers from overfitting the training data, we pre-selected a subset of 10 electrodes to roughly cover the scalp (see Figure 2.3 in Chapter 2 on page 43). Also, to reduce the number of temporal features per epoch, we binned the signal in 100 ms long time-bins. Thus, in total, there were 120 features (10 electrodes $\times$ 12 time-bins) for each trial. We used two, arguably the most simple, linear classifiers: linear discriminant analysis (LDA; Fisher, 1936) and support vector machine (SVM; Cortes & Vapnik, 1995). To reduce chances of overfitting, both LDA and SVM models were regularized, which means that the models were penalized if they were too complex. The regularization parameters for both LDA and SVM were set at 0.5.

Classifiers were trained and tested with stratified 10-fold cross-validation, except when classifying between correct rejections and false alarms, in which case, we used 5-fold stratified cross-validation to help reduce instances of training or test folds containing examples from one condition only. A small number of participants ($N = 3$) had to be excluded when classifying between correct rejections and false alarms, for they had one or more test folds containing examples from one class only.

The cross-validation folds were stratified, which means that the ratio of the number of trials for the two conditions (or classes) were similar across the training and test folds. Classifier performance was evaluated for the test folds— for each trial in a test fold, the classifier produced a score, which could be transformed into the posterior probability for the trial to belong in one class over the other. The scores for all trials in the test fold were used to plot the ROC, and to calculate AUC of the ROC. AUCs were averaged across all the test folds for a participant.

Class imbalance When classifying between hits and misses, or correct rejections and false alarms, total number of trials for the two classes were imbalanced; participants made more hits than misses, more correct rejections than false alarms. In these situations, it is possible for the classifier to be biased towards predicting the over-represented class (hits and correct rejections, respectively). To prevent that, we used over-sampling of the number of trials for the under-represented class (misses and false alarms, respectively). We used Synthetic Minority Oversampling technique (SMOTE; Chawla et al., 2002; Arora et al., 2018; Chakravarty et al., 2020), which creates additional examples for the under-represented class using a nearest-neighbour approach. Importantly, oversampling was done within the cross-validation procedure, only for the under-represented class in the training-folds. Others pursuing similar problems have opted for under-sampling of over-represented class more commonly (e.g., Noh et al., 2014; Watanabe et al., 2011). However, for the current analyses, the total number of trials were small, and thus, under-sampling may also lead to poor training of the classifier, simply due to the lack of enough examples. Also, in light of the results in Chapter 2 (Chakravarty et al., 2020), we suspected that classifier performance, across participants, may not be significantly different with and without oversampling by SMOTE, probably because overall, the predictions were small.

Analysis of feature-importance with LDA weights The multivariate classifiers estimate pattern of activity that best discriminates between the two classes. In this sense, the classifiers work similar to how the brain processes information— by analyzing patterns of neural activity. However, it is possible for the classifier-identified pattern of brain activity to be different from the pattern of activity used by the brain to produce a specific behaviour. With non-linear classifier models, including the state-of-the-art neural network models, it is not straightforward to examine the classifier-identified patterns of brain activity, in order to compare them with knowledge obtained from the planned-comparisons and descriptive methods. However, with simple linear classifiers like LDA, it is straightforward to look into classifier-identified patterns of brain activity. Specifically, the coefficient of each feature in the training set of an LDA model reflects its importance or weight with respect to other features. We used the LDA weights to determine which spatio-temporal features were deemed more important by LDA for each classification problem. For each participant, the weights were first averaged across all training folds, then re-scaled to the interval $[0, 1]$. Finally the weights were averaged across participants. Then we looked into the changes in the averaged weights with time, and also their distribution across the scalp. We considered only those participants for whom LDA achieved better-than-chance predictions.

Excluding influence of motor-preparatory activity with vincentization Since test-related activity also coincides with the action of making responses, it is possible for the classifiers to pick up on differences due to motor-preparatory activity rather than memory-relevant brain activity. For example, in this data set, hits were faster than misses (see Table 3.1)— the classifier may be use this difference to make the predictions for hits and misses. To check this, we conducted follow-up analysis with vincentized time bins (Ratcliff, 1979; Weidemann & Kahana, 2019a), which can account for the potential influence due to motor-preparatory activity.

First, for each test trial, we used the EEG signal starting from stimulus-onset and up to the response (signal was truncated 40 ms advance of the response). The rest of the signal was excluded. Notably, in this case, we did not stick to our default 1200 ms time-window of analysis and instead for longer response times, epochs were also captured for a longer time-interval. Then, for the classifier analysis, to equate the number of features across trials, the number of time-related features (mean amplitudes over specific time-intervals)

were calculated by taking weighted averages— to produce a total of 12 time-related features, as before. Thus, for a trial with response time of 1200 ms, we would obtain 12 time-related features, each averaged over 100 ms long time-intervals; for a trial with response time of 600 ms, we would also obtain 12 time-related features but each of them is averaged over 50 ms long time-intervals. All time-intervals are aligned to the onset of the test stimulus (0 ms).

Classifier performance as functions of time To investigate how memory-relevant information evolved over time, we conducted moving-window based classification analyses, following the approach of Weidemann and Kahana (2019a). First, we extracted the EEG signal for individual 100 ms time windows starting from 0 to 1200 ms post stimulus-onset. Then, we trained and tested the classifiers for each of these 100 ms long signals. We refer to this as the independent time-bin analysis. Next, we trained and tested the classifiers by sequentially adding the 100 ms long signals to it. This means that while the first set of classifiers were trained and tested on the 0–100 ms signal, the second set of classifiers were trained and tested on the 0–200 ms signal and so on. We shall call this the cumulative time-bin analysis. Both set of analyses were done with and without vincentization to account for the influence due to motor-preparatory activity, as described above.

All analyses were done in MATLAB (2018b) along with specific functions from the Statistics and Machine Learning Toolbox (Martinez et al., 2017). Two-tailed t-tests against chance (0.5) were used to examine classifier success across participants; significant effects were relative to $\alpha = 0.05$. Bayes Factors, obtained from Bayesian t-tests are also reported. The Bayesian t-tests were carried out using a function written for MATLAB by SamPenDu (2015). The Bayes Factor (BF_{10}) indicates the Bayesian probability for the alternative over the null hypothesis; $BF_{10} = \frac{p(H_1)}{p(H_0)}$. $BF_{10} > 10$ indicates strong evidence for the alternate and $BF_{10} < 0.1$ indicates strong evidence for the null (Kass & Raftery, 1995). $BF_{10} > 3$ indicates moderate evidence for the alternate and $BF_{10} < 0.3$ indicates moderate evidence for the null. For the one-sample t-tests, we also report the 95% confidence intervals of the mean for the classifier success; classifier success was measured by the AUC of the ROC. The 95% CIs were used as an estimate of the size of the classifier success across participants. Also, for reproducibility of the classification results, a pseudo-random number generator was used (Mersenne twister, seed = 0).

Condition	Accuracy (%)	Response time (ms)
Hits (old)	77.6 (9.93)	978 (195)
Misses (old)	21.74 (9.84)	1361 (446)
Correct rejections (new)	85.96 (12.07)	1105 (273)
False Alarms (new)	13.41 (11.80)	1568 (564)

Table 3.1: Mean accuracy and response times for the different memory outcomes. Standard deviations are in parentheses next to the mean values.

3.3 Results

3.3.1 Behaviour

Table 3.1 presents average accuracy and response times for hits, misses, correct rejections and false alarms. Hits and correct rejections were more frequent than misses or false alarms. Also, hits were faster than misses. The standard deviations (reported in parentheses in Table 3.1) for the response times were large, suggesting meaningful variability across participants.

3.3.2 ERPs at test

Traditional ERP analysis for this dataset was previously reported in Chen et al. (2014) and showed significant ERP effects for the FN400 and LPP and for both the old versus new, and the hit versus miss contrasts. However, as described in the Methods, to better identify artifactual trials, here we used an absolute voltage threshold and a point-to-point difference threshold (also see Chapter 2; Chakravarty et al., 2020), calling for a re-analysis of the ERPs at test to check if the old/new and retrieval-success effects were still significant. Figure 3.1 presents grand averaged ERPs at test, for the frontal electrode Fz (Figure 3.1a,c) and the left parietal electrode P3 (Figure 3.1b,d), and compares between hits and correct rejections (Figure 3.1a,b) and hits and misses (Figure 3.1c,d). Consistent with Chen et al. (2014), at Fz, mean amplitude over the 300–500 ms time window, post stimulus-onset, was significantly more positive for hits than correct rejections, $[t(60) = 3.63,\ p < 0.001,\ BF_{10} = 42.66]$, or for misses, $[t(60) = 4.94,\ p < 0.001,\ BF_{10} > 100]$. The Bayes Factors strongly supported the effect in both cases ($BF_{10} > 10$). Likewise, at P3, mean amplitude over the time window 500–800 ms, post stimulus-onset, was significantly more positive for hits than correct rejections $[t(60) = 5.79,\ p < 0.001,\ BF_{10} > 100]$ or for misses $[t(60) = 3.57,\ p < 0.001,\ BF_{10} = 36.67]$. Again, the The Bayes Factors strongly supported the effect in both cases ($BF_{10} > 10$). The

respective topographic plots of the difference waves (hits − correct rejections and hits − misses) also confirmed a frontal positive signal for FN400 and a posterior positive signal for LPP. Thus, the old/new and retrieval-success ERP effects were still present. We proceed with the predictive analysis as planned.

3.3.3 Predictions with univariate measures of the FN400 and LPP

First, we tested if FN400 and LPP amplitude predicted memory for individual trials, across participants (see Figure 3.2 and Table 3.2). FN400 amplitude achieved significant success in classifying old and new trials, but the Bayes Factor was inconclusive for this classification ($BF_{10} = 1.12$). FN400 amplitude classified hits and misses and perceived-old and new trials with significant success, and with the Bayes Factors strongly supporting the effects ($BF_{10} > 10^2$). However, FN400 amplitude did not classify between correct rejections and false alarms. Interestingly, the 95% CI of the FN400-based AUCs for classifying hits and misses, [0.53 0.55], was greater than that for old and new trials, [0.50 0.52], and also greater than that for perceived-old and new trials, [0.51 0.53], suggesting that FN400 amplitude may be more specific to the difference due to memory-success.

LPP amplitude was also successful at classifying between old and new trials (with $BF_{10} > 10^2$), hits and misses, and perceived-old and new trials; and also failed to classify between correct rejections and false alarms. However, unlike FN400, the 95% CI of the LPP-based AUCs for classifying hits and misses, [0.52 0.55], largely overlapped with that for classifying old and new trials, [0.52 0.54], and was the same for the perceived-old- and new trials, [0.52 0.55], suggesting that LPP amplitude, when compared to FN400 amplitude, may be less specific to difference due to memory-success, and more specific to difference due to targets and lures. Also, LPP amplitude predicted significantly better than FN400 amplitude for classifying old and new trials ($t(60) = 4.09,\ p < 0.001,\ BF_{10} > 10^2$), with the Bayes Factor strongly supporting the effect. This difference between predictions with LPP and FN400 amplitudes was also significant for classifying perceived old and new trials ($t(60) = 2.08,\ p < 0.05,\ BF_{10} = 1.05$), but the Bayes Factor was inconclusive. However, there was no significant difference between predictions with LPP and FN400 amplitudes for classifying hits and misses ($t(60) = 0.51,\ p = 0.61,\ BF_{10} = 0.16$), though the Bayes Factor was inconclusive. Taken together, these differences between predictions with LPP and FN400 amplitudes may further support the suggestion that FN400 was more specific to difference

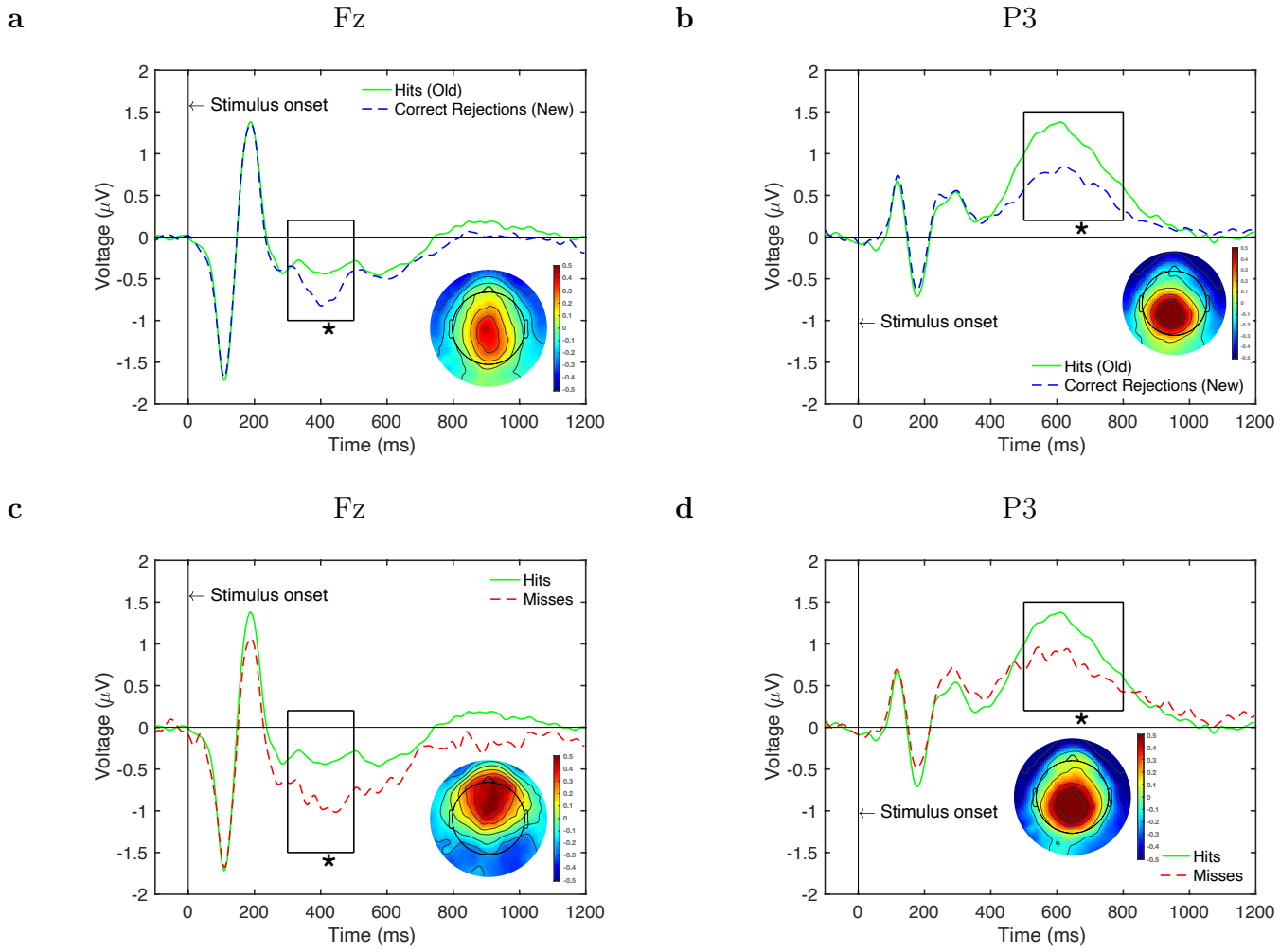

Figure 3.1: Grand averaged ERPs at test, comparing hits and correct rejections (upper panels) and hits and misses (lower panels). ERPs are plotted separately for the frontal electrode Fz (a,c) and the left parietal electrode P3 (b,d) to examine the effects of the FN400 and the LPP respectively. Corresponding topographic maps are plotted for the difference waves (hits − CR or hits − misses) for the window of the FN400 (a,c) or the LPP (b,d) respectively; color indicates mean voltage (μV).

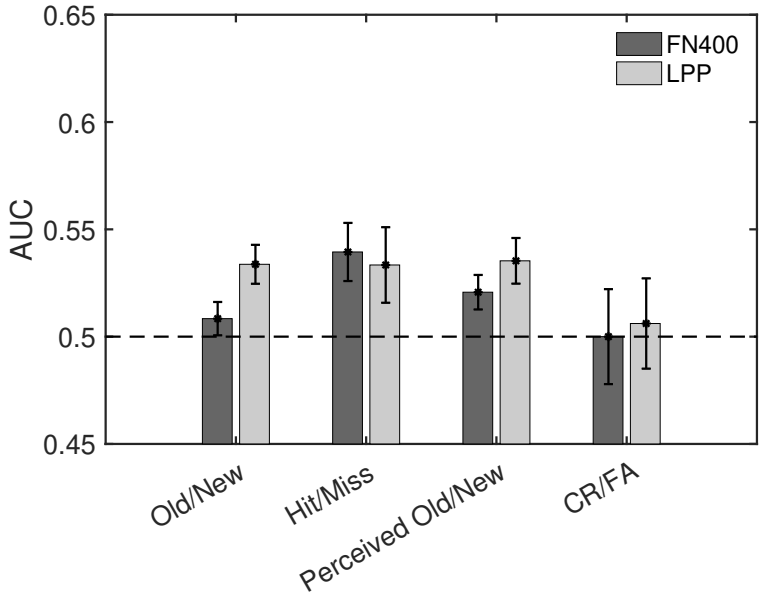

Figure 3.2: Classifications with FN400 (computed from electrode Fz) and LPP (computed from electrode P3) amplitudes, separately for the classifications between 1) old and new trials, 2) hits and misses, 3) perceived-old- and new trials, and 4) correct rejections and false alarms. Error bars are 95% confidence intervals. Dashed black line refers to chance (0.5).

due to memory-success than LPP.

In sum, predictions with both FN400 and LPP amplitudes achieved significant— but small— success in classifying different memory outcomes. However, neither FN400 nor LPP could predict correct rejections and false alarms.

To better understand the relation between predictions with FN400 and LPP amplitudes, as well as the relation between predictions for the four classification problems, we analyzed the correlations between AUCs obtained with the different measures (and classification problems) across participants, as we discuss below.

FN400	Old/New*	$t(60) = 2.12,\ p < 0.05,\ BF_{10} = 1.12,\ 95\%\ CI = [0.50\ \ 0.52]$
	Hit/Miss*	$t(60) = 5.71,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.53\ \ 0.55]$
	Perceived Old/New*	$t(60) = 5.05,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.51\ \ 0.53]$
	CR/FA	$t(60) < 0.01,\ p = 0.99,\ BF_{10} = 0.14,\ 95\%\ CI = [0.48\ \ 0.52]$
LPP	Old/New*	$t(60) = 7.28,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.52\ \ 0.54]$
	Hit/Miss*	$t(60) = 3.72,\ p < 0.001,\ BF_{10} = 55.63,\ 95\%\ CI = [0.52\ \ 0.55]$
	Perceived Old/New*	$t(60) = 6.51,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.52\ \ 0.55]$
	CR/FA	$t(60) = 0.57,\ p = 0.57,\ BF_{10} = 0.16,\ 95\%\ CI = [0.48\ \ 0.53]$

Table 3.2: Classifications with FN400 and LPP amplitudes, t-test against chance (0.5) for the AUCs, along with the Bayes Factor (BF_{10}). Significant effects are marked with *.

Notably, Chen et al. (2014) found that across participants, the difference in FN400 amplitude for hits and correct rejections (or the old/new effect) did not correlate to the same difference computed for the LPP amplitude. Likewise, the difference between hits and misses (or the retrieval-success effect), as captured by the FN400 and LPP amplitudes, also did not correlate with each other. In other words, although both FN400 and LPP showed significant difference due to the old/new and retrieval-success effects, the effects were not correlated for the two ERPs, suggesting that FN400 and LPP may not be sensitive to common variability in memory retrieval processes.

Supporting this idea, here we found no correlation between the predictive performances (AUCs) of FN400 and LPP amplitudes, for any of the four classification problems— old and new trials, hits and misses, perceived-old- and new trials, and correct rejections and false alarms (see Figure 3.3). Analysis of partial correlations also showed that the AUCs obtained with FN400 and LPP amplitudes did not relate to each other, for any of the classification problems.

We also conducted partial correlations for the AUCs obtained with the FN400 amplitudes, across the four classification problems, while controlling for the effects of predictions (AUCs) based on LPP amplitudes (also for the four classifications). This showed significant positive correlations between 1) classification of old and new trials, and classification of perceived-old- and new trials ($p < 0.001$) and 2) classification of hits and misses, and classification of perceived-old- and new trials ($p < 0.001$); all other correlations were non-significant. Since classification of hits and misses correlated with the classification of perceived-old- and new trials, but not with the classification of old and new trials, it could suggest that FN400 amplitude is more specific to response-related differences than more general differences due to targets and lures. However, the classification of old versus new trials with FN400 amplitude was deemed inconclusive by the Bayes Factor, which may undermine the above interpretation.

Partial correlations for the AUCs obtained with the LPP amplitudes, across the four classification problems, while controlling for the effects of predictions (AUCs) based on FN400 amplitudes (also for the four classifications), showed significant correlations among classifications of old and new trials, hits and misses, and perceived-old- and new trials (pair-wise, $p < 0.001$). Thus, apart from the classification of correct rejections and false alarms, which had failed with both FN400 and LPP amplitudes, in case of LPP, the three other classifi-

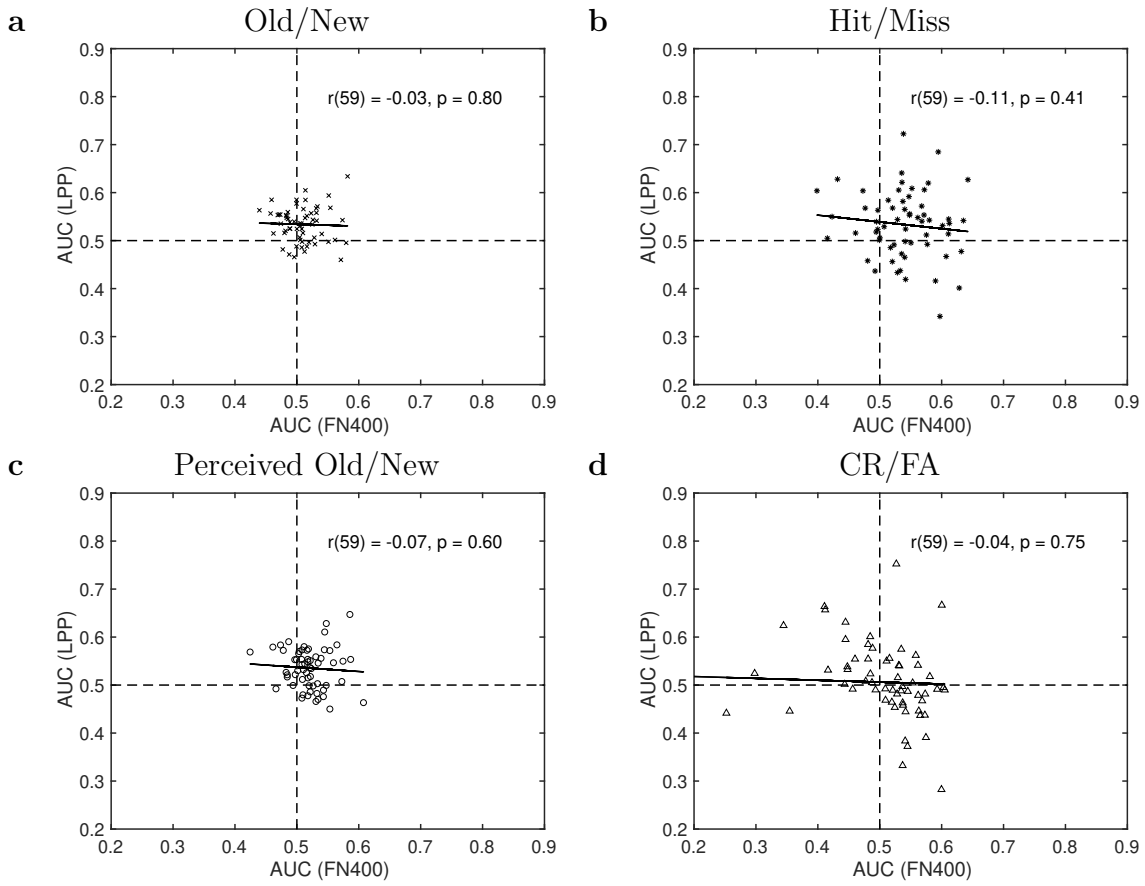

Figure 3.3: Correlation between AUCs obtained from FN400 and LPP based classifications, for each classification problem. Dashed black lines refer to chance (0.5).

cations agreed with each other; in case of FN400, both classifications of old and new trials, and hits and misses agreed with the classification of perceived-old- and new trials.

Chen et al. (2014) also found that for the retrieval-success effect, participants' d' and average response-times (for hits only), were significantly correlated with the FN400 amplitude. However, LPP amplitude was not correlated with d' or response-times, either for the old/new- or the retrieval-success effect. Paralleling their results, when classifying hits and misses, we found a significant positive correlation between d' and AUCs obtained with the FN400 amplitudes (see Figure 3.4), and a trend ($p = 0.05$) for a negative correlation between average response-times (hits) and the FN400-based AUCs (Figure 3.5). Interestingly, FN400-based AUCs correlated negatively with d' and response-times for classifying correct rejections and false alarms. The negative correlation with d' could be because for smaller d', false alarms were more likely, which could have helped better discriminate false alarms from correct rejections than when false alarms were very rare, as in the case of a higher d'. On the other hand, LPP-based AUCs correlated with d' when classifying old and new trials, and perceived-old- and new trials; the correlation between LPP-based AUCs and response-times, when classifying old and new trials, approached significance ($p = 0.05$). All other correlations were non-significant. Together, these results may provide more support for FN400 being more relevant to difference due to memory-success in this task, whereas the LPP was sensitive to difference due to targets and lures.

3.3.4 Shorter response times

In general, test-related activity can overlap with motor preparatory activity for making responses. We wondered if classifications with FN400 and LPP amplitudes were influenced by motor-preparatory activity. Motor-preparatory activity in the signal could have indexed memory outcomes in an artifactual way. For example, classifying hits and misses could operate on the difference in response times for hits and misses (faster for hits, slower for misses). Such classification of hits and misses is not based on memory-relevant brain activity. For our default choice of epoch-length (1200 ms), trials with shorter response-times (<1200 ms) will likely include motor-preparatory activity in the epoched signal. To check this, we conducted a follow-up analysis by excluding shorter response-times trials. For classifications with FN400 (latency: 300–500 ms post stimulus-onset), we excluded trials with response-times <500 ms. For LPP (latency: 500–800 ms post stimulus onset), we excluded trials with

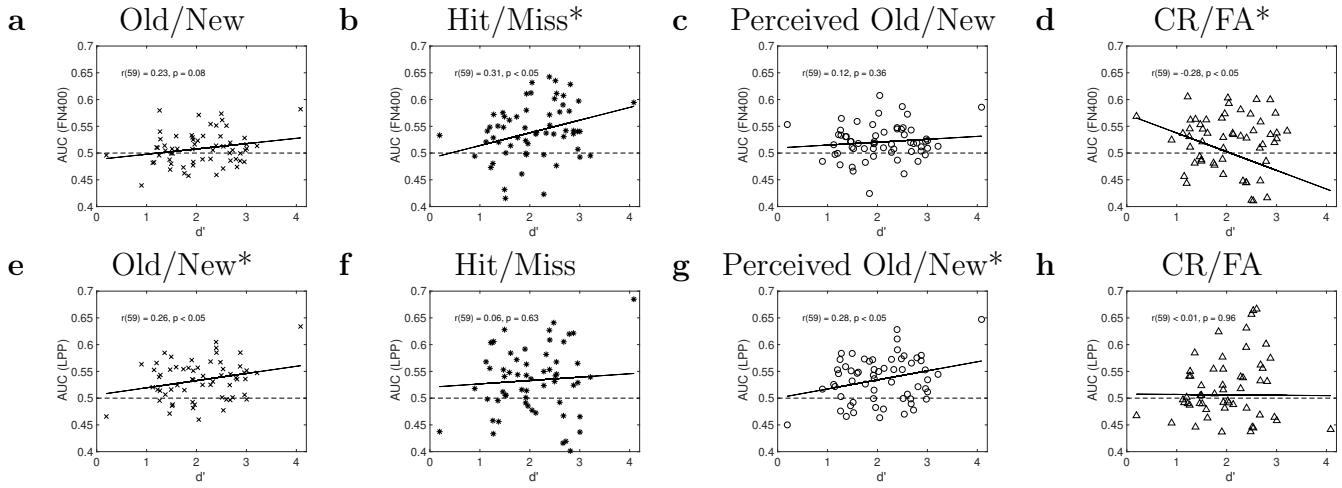

Figure 3.4: Correlation between AUCs obtained from FN400 or LPP amplitude-based classifications and d' values of participants behaviour, separately for each classification problem. Dashed black lines refer to chance (0.5). Panels with significant correlations are marked with *.

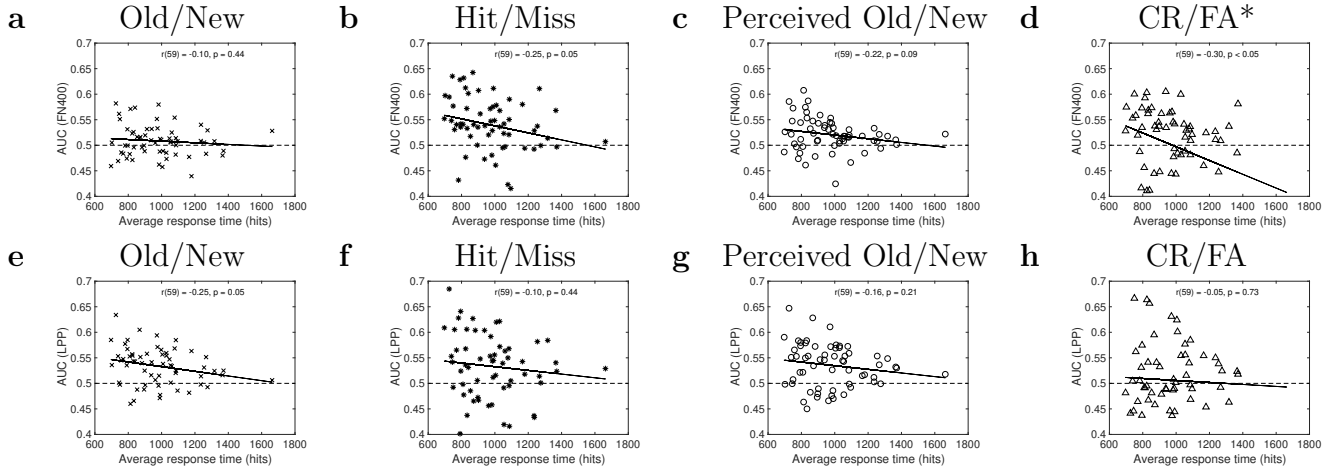

Figure 3.5: Correlation between AUCs obtained from FN400 or LPP amplitude-based classifications and average response-times (for hits only), separately for each classification problem. Dashed black lines refer to chance (0.5). Panels with significant correlations are marked with *.

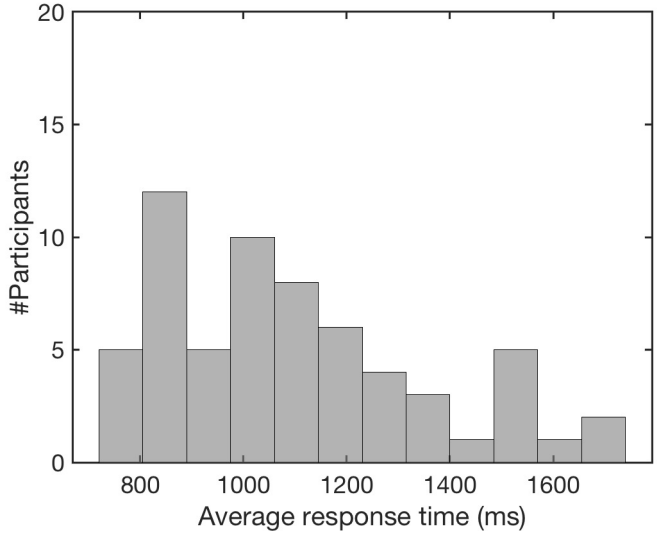

Figure 3.6: Average response times across participants.

response-times <800 ms.

For FN400, classification results were similar to that without exclusion of shorter response-time trials (see Figure 3.7 and Table 3.3). Thus, results based on the FN400 amplitude, as presented above, were not influenced by motor-preparatory activity. However, with the above-mentioned exclusion-criterion for FN400 (exclude trials with response times shorter than 500 ms), we did not end up rejecting many trials, because the average response-time, across all types of trials, exceeded 500 ms (see Figure 3.6).

For LPP, classification of hits and misses became non-significant after excluding shorter response-time trials. Classifications of old and new trials, and perceived-old- and new trials, were significant, but of smaller size (indicated by the 95% CIs) than without the exclusion of trials. Thus, predictions based on LPP amplitudes, for the classification of old and new trials, or the classification of perceived-old- and new trials, were no longer better than predictions based on FN400 amplitudes, as we had found before. Interestingly, classification of correct rejections and false alarms with the LPP amplitudes, after the exclusion of shorter response-time trials, became significant (see Table 3.3). Since correct rejections were, on average, faster than false alarms; and false alarms were more infrequent than correct rejections, correct rejection trials may have been excluded more often, which could have led to better discrimination between the two types of trials.

Thus, when shorter response-time trials, which are less likely to include the LPP, were

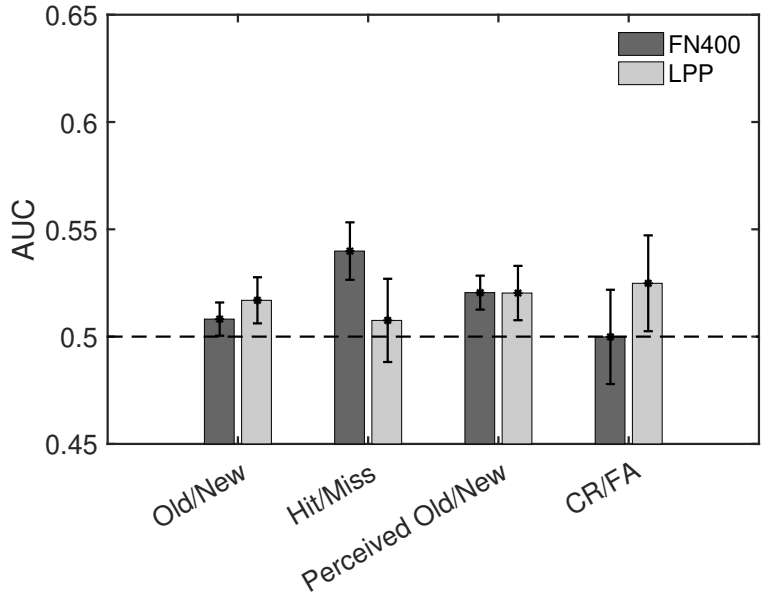

Figure 3.7: Classification based on the FN400 (computed from electrode Fz) and the LPP (computed from electrode P3) after rejecting trials with response times lesser than 500 ms and 800 ms respectively for the FN400 and the LPP. Results are grouped into four different classification problems: 1) old versus new, 2) hits versus misses, 3) perceived old versus perceived new and 4) correct rejections versus false alarms. Error bars are 95% confidence intervals. Dashed black line refers to chance (0.5).

FN400	Old/New*	$t(60) = 2.07,\ p < 0.05,\ BF_{10} = 1.01,\ 95\%\ CI = [0.50\ \ 0.52]$
	Hit/Miss*	$t(60) = 5.83,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.53\ \ 0.55]$
	Perceived Old/New*	$t(60) = 5.08,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.51\ \ 0.53]$
	CR/FA	$t(60) < 0.01,\ p = 0.99,\ BF_{10} = 0.14,\ 95\%\ CI = [0.48\ \ 0.52]$
LPP	Old/New*	$t(60) = 3.09,\ p < 0.005,\ BF_{10} = 9.79,\ 95\%\ CI = [0.51\ \ 0.53]$
	Hit/Miss	$t(60) = 0.76,\ p = 0.45,\ BF_{10} = 0.19,\ 95\%\ CI = [0.49\ \ 0.53]$
	Perceived Old/New*	$t(60) = 3.14,\ p < 0.005,\ BF_{10} = 11.37,\ 95\%\ CI = [0.51\ \ 0.53]$
	CR/FA*	$t(60) = 2.18,\ p < 0.05,\ BF_{10} = 1.26,\ 95\%\ CI = [0.50\ \ 0.55]$

Table 3.3: Predictions based on FN400 and LPP after excluding trials with shorter response times, t-test against chance (0.5) for the AUCs. Significant effects are marked with *.

excluded, predictions based on LPP amplitudes became overall worse. Notably, with the exclusion criterion for LPP (response times shorter than 800 ms) we ended up rejecting many trials— on average 43.25% trials were rejected for the old/new or the perceived old/new; 47.50% for the hit/miss and 38.99% for the correct rejection/false alarm classifications, confounding the interpretation of results for LPP. Alternatively, the above results for LPP may be because recollection is more likely to occur following a quick judgement, that can be based on familiarity.

3.3.5 Predictions with multivariate EEG activity at test

Next, we examined the multivariate spatio-temporal EEG signal during the test phase to predict the different memory outcomes, using the two classifiers: LDA and SVM. Assuming that FN400 and LPP amplitudes index different cognitive processes behind recognition-memory outcomes; and also considering that there may even exist other signals during test-phase activity that is also relevant to memory (i.e., beyond FN400 and LPP), we expected to find better predictions with the multivariate methods, which analyzed patterns of activity.

We tested if LDA and SVM classified between old and new trials, hits and misses, perceived-old- and new trials, and correct rejections and false alarms, for the multivariate EEG signal at test (Figure 3.9 and Table 3.4). Both classifiers achieved significant success, and for all four classifications, the Bayes Factor strongly supported each of the effects. Across participants, predictions with LDA and SVM were very similar to each other (see Figure 3.8), suggesting generalizability of the results across the two chosen classifier models. Classification of old and new trials, with LDA and SVM, produced a 95% CI of [0.56 0.59], which was greater than, and non-overlapping with the 95% CI for classifying old and new trials with FN400 amplitudes, [0.50 0.52] or LPP amplitudes, [0.52 0.54]. The 95% CIs for LDA and SVM were also greater than that for FN400 or LPP amplitude-based classification of hits and misses, and classification of perceived-old and new trials. Further, unlike FN400 or LPP amplitude-based predictions, both LDA and SVM achieved significant success in classifying correct rejections false alarms, and with good margins (95% CI for LDA: [0.54 0.58]; SVM: [0.53 0.58]). Overall, the multivariate classifiers performed substantially better than FN400 or LPP amplitudes-based predictions alone.

Similar to the correlations between predictions with FN400 and LPP amplitudes, here too, we looked into the relations between the four classifications, in terms of the perfor-

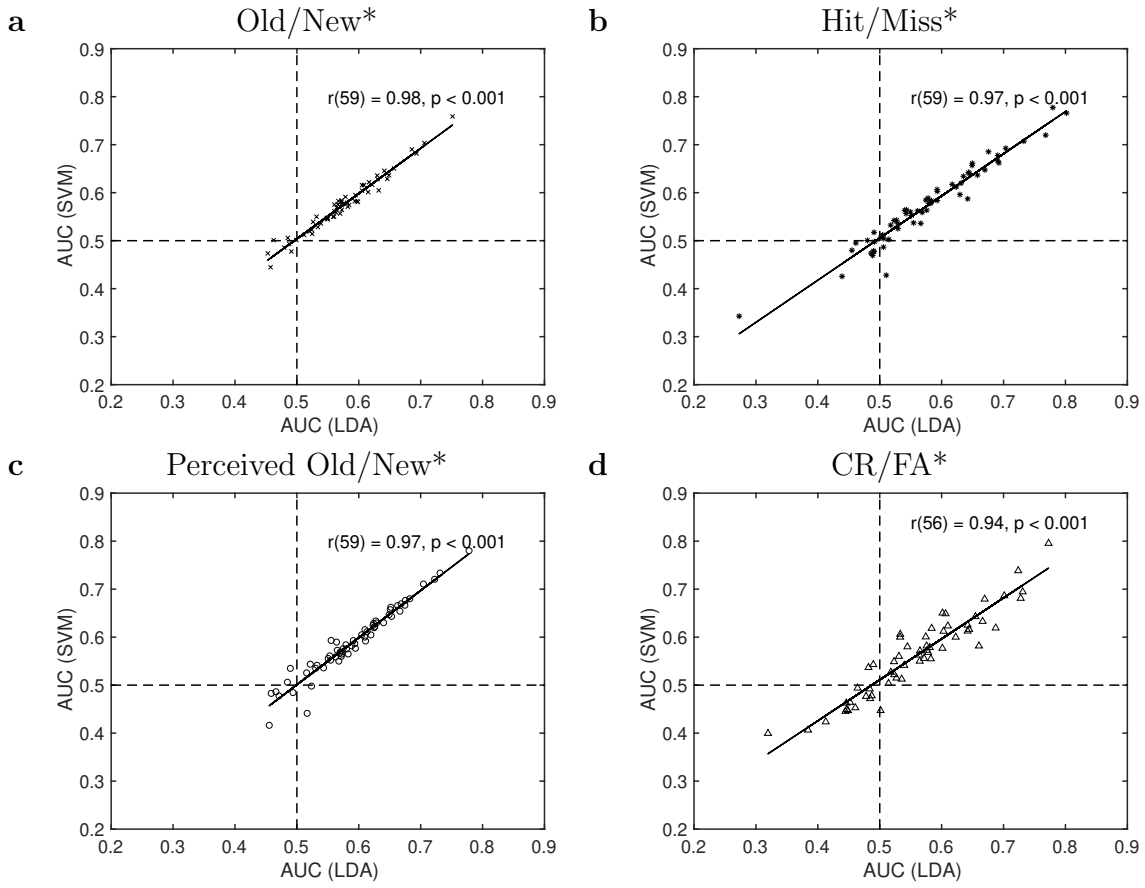

Figure 3.8: Correlation between LDA and SVM classifiers, separately for each classification problem. Dashed black lines refer to chance (0.5). Panels with significant correlations are marked with *.

LDA	Old/New*	$t(60) = 9.98,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.56\ \ 0.59]$
	Hit/Miss*	$t(60) = 6.80,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.55\ \ 0.59]$
	Perceived Old/New*	$t(60) = 10.09,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.57\ \ 0.61]$
	CR/FA*	$t(57) = 5.53,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.54\ \ 0.58]$
SVM	Old/New*	$t(60) = 9.63,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.56\ \ 0.59]$
	Hit/Miss*	$t(60) = 6.47,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.55\ \ 0.60]$
	Perceived Old/New*	$t(60) = 10.30,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.57\ \ 0.61]$
	CR/FA*	$t(57) = 4.82,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.53\ \ 0.58]$

Table 3.4: Multivariate classification with LDA and SVM, t-test against chance (0.5) for the AUCs, along with Bayes Factors (BF_{10}). Significant effects are marked with *.

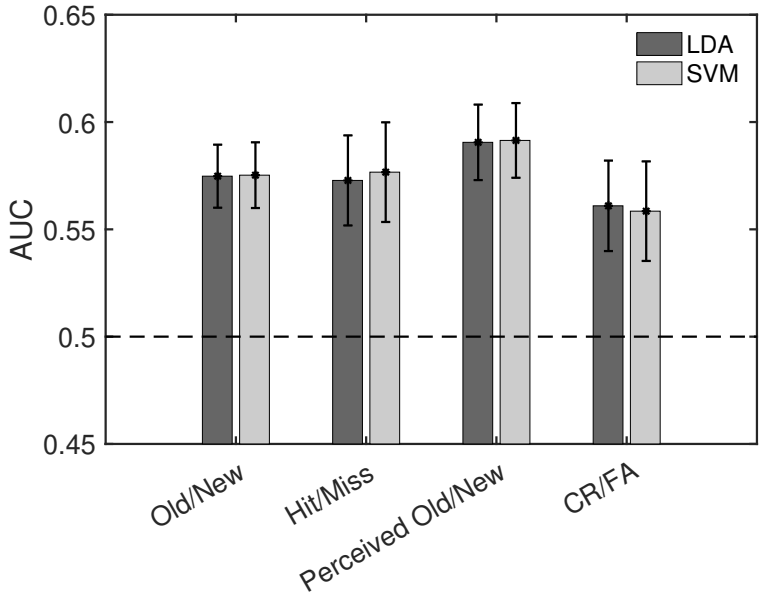

Figure 3.9: Multivariate classification with LDA and SVM. Results are grouped into four classification problems: 1) old and new trials, 2) hits and misses, 3) perceived-old- and new trials, and 4) correct rejections and false alarms. Error bars are 95% confidence intervals. Dashed black line refers to chance (0.5).

mance of two classifiers. Partial correlations for LDA AUCs, across the four classifications showed that classification of perceived-old- and new trials was correlated positively with 1) classification of old and new trials [$r(56) = 0.65,\ p < 0.001$], 2) classification of hits and misses [$r(56) = 0.47,\ p < 0.001$] and 3) classification of correct rejections and false alarms [$r(56) = 0.27,\ p < 0.05$]. Partial correlations for SVM AUCs, across the four classifications also showed significant positive correlations between 1) classification of perceived-old- and new trials and classification of old and new trials [$r(56) = 0.65,\ p < 0.001$], 2) classification of perceived-old- and new trials and classification of hits and misses, [$r(56) = 0.40,\ p < 0.005$]; and 3) a trend for classification of perceived-old- and new trials and classification of correct rejections and false alarms, [$r(56) = 0.23,\ p = 0.09$]. Overall, there was underlying similarity between the classification of perceived-old- and new trials and the other three classifications, for both LDA and SVM.

Chapter 2 (Chakravarty et al., 2020) reported meaningful variability in classifier performance (for the multivariate EEG signal during the study phase) as a function of participant's d'. If the participant performed better in the task, the classifier also tended to achieve a higher AUC. Here too, we found significant positive correlations between d' and AUCs of

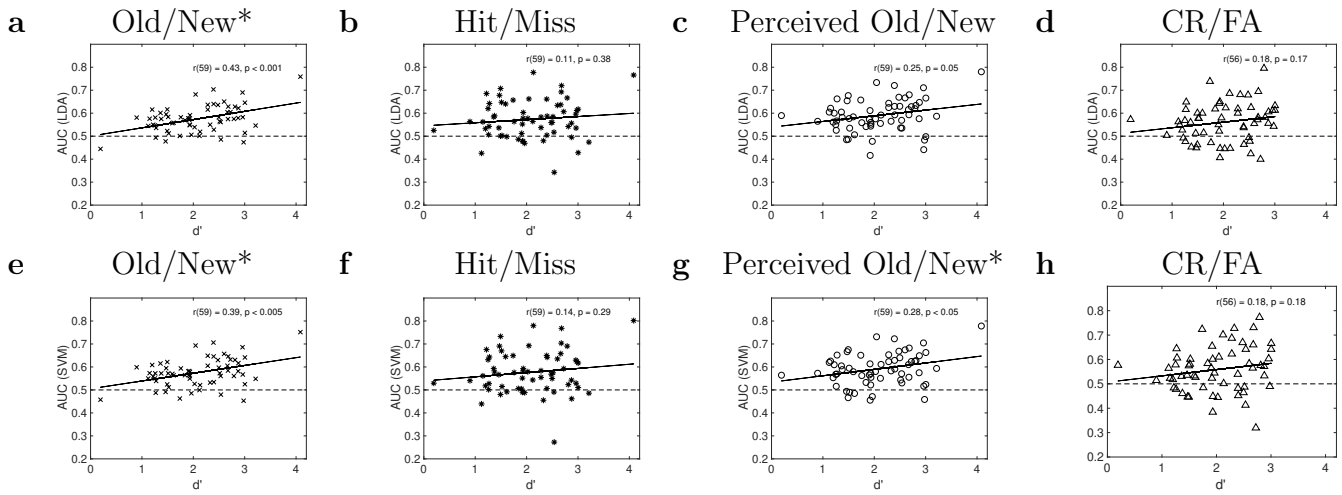

Figure 3.10: Correlation between LDA and SVM AUCs and participant's d', separately for each classification problem. Dashed black lines refer to chance (0.5). Panels with significant correlations are marked with *.

the LDA and SVM classifiers, for classifying old and new trials (see Figure 3.10). SVM also correlated significantly with d' for classifying perceived-old- and new trials, whereas LDA showed a trend effect for the same ($p = 0.05$). Taken together, these correlations were consistent with the suggestion previously made in Chapter 2 (Chakravarty et al., 2020), that for better-performing participants, the corresponding brain activity may be more task-relevant, which is picked up by the classifiers, leading to better predictions.

3.3.6 Analysis of LDA feature weights

Since FN400 and LPP amplitudes failed to classify between correct rejections and false alarms, whereas both LDA and SVM succeeded for this classification, and also since LDA- and SVM-based predictions were overall better than FN400 and LPP amplitude-based predictions, we wondered which features the classifiers may have relied more on, to distinguish correct rejections from false alarms. As mentioned in the Methods, with LDA, the coefficient for each feature (in the training set) indexes the relative importance or weight of the feature in comparison to other features. Accordingly, to understand the features of relative importance to the LDA classifier, for each classification problem, we looked into the spatial and temporal distribution of the LDA weights, averaged across participants with AUCs > 0.5 (or the successful cases).

LDA feature-weights, as functions of the 12 time-related features or the time-bins (Fig-

ure 3.11), showed very similar trends for classification of old and new trials, and classification of perceived-old- and new trials— relatively smaller weights for the earlier time-bins, followed by a peak for the 6^{th} time-bin (which included the signal averaged over 500–600 ms, post stimulus-onset) and thereafter, a drop in the weights for the later time-bins. In other words, the averaged signal over the 500–600 ms time-interval was deemed most important by LDA, for classifying both old and new trials, and perceived-old- and new trials. Considering the latency of LPP (500–800 ms post stimulus onset), the peak around the 6^{th} time-bin may suggest a greater influence of LPP, for classifying the old and new trials, and the perceived-old- and new trials.

LDA weights, for classifying hits and misses, showed an earlier smaller peak (for the first two time-bins), followed by a second peak around the 5^{th} and 6^{th} time-bins; this was followed by a drop in weights for the later time-bins. Thus, for classifying hits and misses, LDA followed a different pattern than that for classifying old and new trials, or for classifying perceived-old- and new trials. For classifying hits and misses, LDA assigned greater importance to more than one signal, which may potentially index different memory-relevant cognitive processes.

Finally, for classifying correct rejections and false alarms, the LDA weights curve was less undulating— LDA weights for the earlier time-bins (1^{st} to 6^{th}) were overall greater than LDA weights for the later time-bins. There was no clear peak around the middle time-bin, as seen for the other three classifications. Thus, the pattern of activity, as identified by LDA, for classifying correct rejections and false alarms was clearly different from that for the other three classifications. To further examine the characteristics of the LDA weights, next we looked into its distribution across the scalp, separately for each of the 12 time-bins.

LDA feature-weights across the scalp (Figure 3.12) showed that for the 6^{th} time-bin, where a peak was observed in Figure 3.11, the weights were relatively greater for the posterior scalp, for all but the correct rejection versus false alarm classification. Since a posterior scalp topography is also characteristic of the LPP, which also has a similar latency, this could suggest that an LPP-like signal played a greater role in the classifications of old and new trials, hits and misses, and perceived-old- and new trials, but not for the classification of correct rejections and false alarms. Thus, not only did the LDA weights for the classification of correct rejections and false alarms showed an absence of a clear peak for this time-bin (Figure 3.11), the corresponding signal used by LDA in this time-bin may have also been

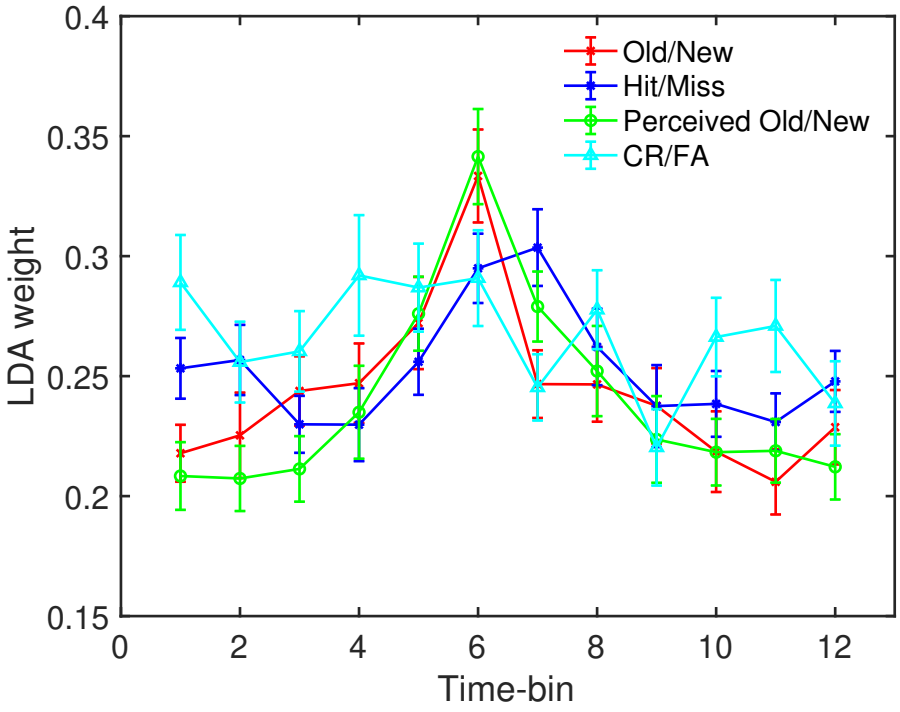

Figure 3.11: Feature-weights for LDA, averaged across participants with LDA AUC > 0.5. Weights are presented for all the time-features (mean amplitude over 100 ms time-intervals) used in the classification analysis, averaged over all included electrodes (the spatial features). Weights are shown separately for the four classification problems. The error bars are standard errors of the mean.

different from an LPP-like signal— the distribution showed greater influence of the bilateral temporal electrodes.

Therefore, we may have failed to see significant success for classifying correct rejections and false alarms with FN400 or LPP amplitudes, because those analyses were restricted to the signal from electrodes Fz and P3. Taken together, this illustrates that there may exist other memory-relevant brain activity that is not captured by the FN400 or LPP ERP effects alone.

Notably, one possible reason behind the peak for the middle time-bins is that average response times (across all trials) were close to 700–800 ms (see Figure 3.6) and thus, the classifiers may have been picking up on motor preparatory activity, as already discussed.

To check this, we conducted a follow-up analysis with the vincentization method, that removed response-related activity from the signal, while keeping same number features across trials for the classification analysis, as we present below.

3.3.7 Classification of the vincentized signal

With vincentization, we obtained a total of 12 time-related features for each trial, for the classification analysis, stopping short of the responses, themselves. We hypothesized that if motor-preparatory activity due to the response actions were picked up by the classifiers and were used to make the decisions, then in this follow-up analysis we would obtain smaller or even non-significant effects for one or more of the classification problems. Alternatively, if motor-preparatory activity, despite being present in the previous set of analysis (with the 1200 ms long epochs), did not influence the classifiers in any substantial way, then these follow-up effects would be similar to that without vincentization.

We found support for the latter— the effects were not significantly different from those without vincentization (see Figure 3.13 and Table 3.5). Thus, even if motor preparatory activity was present, brain activity, starting from the stimulus-onset and prior to making the response, carried information based on which memory outcomes could be classified.

The LDA weights across the 12 vincentized time-bins (averaged across participants with LDA AUC$>$ 0.5) showed an overall different pattern to that without the vincentization (see Figure 3.14)— for all four classifications, there was an earlier peak for the 3^{rd} vincentized time-bin, suggesting greater influence of an earlier signal. For the classification of old and new trials, and classification of hits and misses, there was also a clear second peak for the 5^{th}

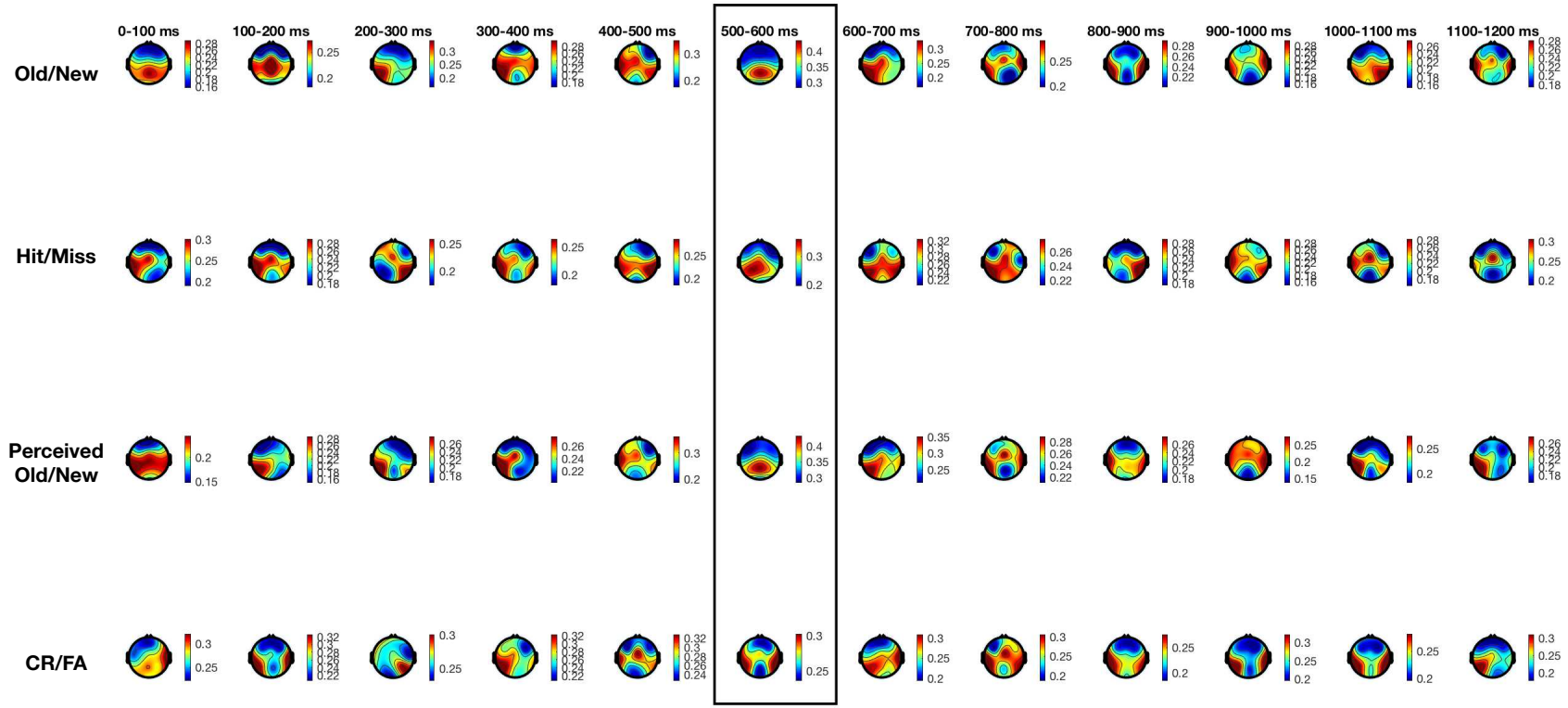

Figure 3.12: Distribution of feature-weights across the scalp, separately for each of the 12 time-intervals, and averaged across participants with LDA AUC > 0.5. The topographic plots were made by interpolating the weights of the electrodes included in the classification analysis to other (not included) electrodes on the scalp, through inverse distance-weighting. Weight-distributions are shown separately for the four classification problems. Colors indicate weights, the color scale varies across panels.

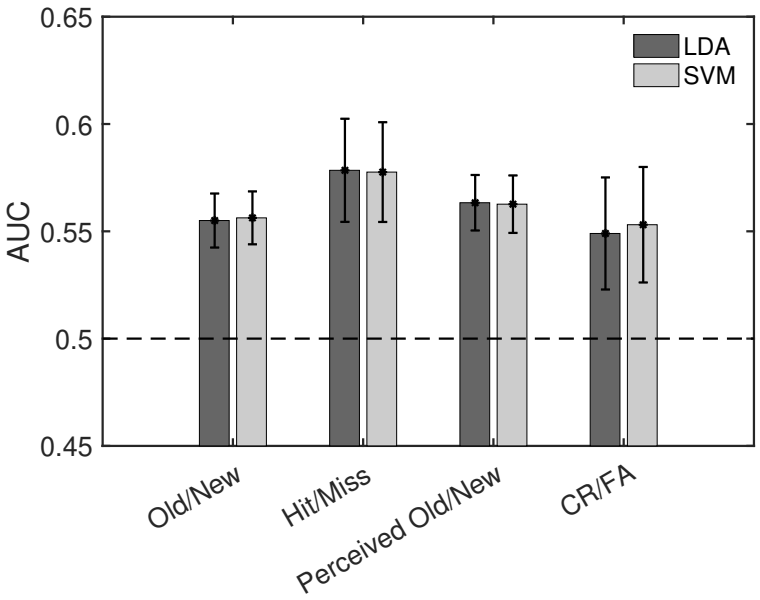

Figure 3.13: Classification with LDA and SVM after truncating the signal for each trial prior to the response. Results are grouped into four different classification problems: 1) old versus new, 2) hits versus misses, 3) perceived old versus perceived new and 4) correct rejections versus false alarms. Error bars are 95% confidence intervals. Dashed black line refers to chance (0.5).

LDA	Old/New*	$t(60) = 8.57,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.54\ \ 0.57]$
	Hit/Miss*	$t(60) = 6.40,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.55\ \ 0.60]$
	Perceived Old/New*	$t(60) = 9.60,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.55\ \ 0.58]$
	CR/FA*	$t(57) = 3.59,\ p < 0.001,\ BF_{10} = 37.57,\ 95\%\ CI = [0.52\ \ 0.58]$
SVM	Old/New*	$t(60) = 8.95,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.54\ \ 0.57]$
	Hit/Miss*	$t(60) = 6.54,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.55\ \ 0.60]$
	Perceived Old/New*	$t(60) = 9.17,\ p < 0.001,\ BF_{10} > 10^2,\ 95\%\ CI = [0.55\ \ 0.58]$
	CR/FA*	$t(57) = 3.77,\ p < 0.001,\ BF_{10} = 63.33,\ 95\%\ CI = [0.52\ \ 0.58]$

Table 3.5: Predictions based on LDA and SVM after truncating the signal for each trial prior to the response, t-test against chance (0.5) for the AUCs. Significant effects are marked with *.

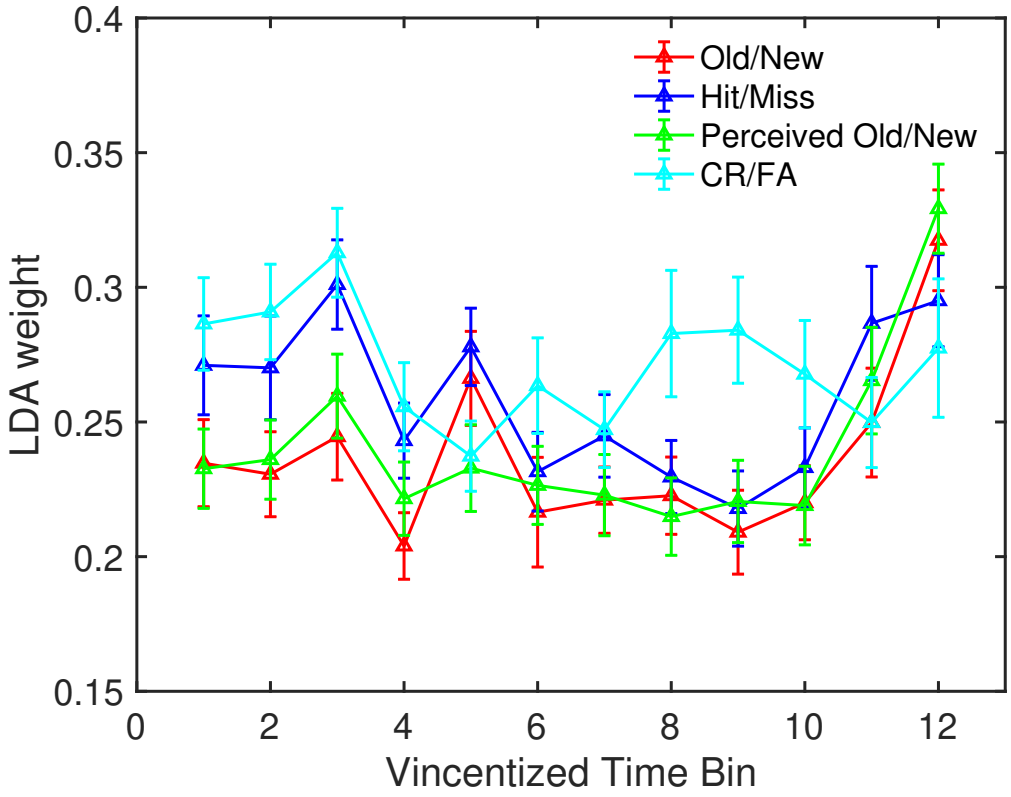

Figure 3.14: LDA feature-weights averaged across participants with LDA AUC > 0.5. Weights are presented across all 12 vincentized time-bins, and averaged over all included electrodes (the spatial features). Weights are shown separately for the four classification problems. The error bars are standard errors of the mean.

time-bin, suggesting that there may be another relevant signal. Thus, for the classification of old and new trials, and the classification of hits and misses, LDA assigned greater importance to multiple signals. For the classification of correct rejections and false alarms, the second peak was at a later (vincentized) time-bin (8^{th}-9^{th}).

Finally, all four LDA weights-curves showed a roughly increasing trend (Figure 3.14), specifically for the later vincentized time-bins, which partially supports the idea that memory-relevant information in the brain may have increased with time, prior to reaching a decision.

3.3.8 Evaluating single- and dual-process accounts with classifier evidence

We adopted the logic introduced by Weidemann and Kahana (2019a), we measured classifier performance as a function of time, we asked if information relevant to making the memory decisions are driven independently by multiple signals, in line with a dual-process account

(Yonelinas, 2002); or if memory-relevant information is supported by a unitary, integrated strength signal, as in a single-process account (Dunn, 2008). For the analysis with the independent time-bins, classifier performance for time-bin t reflected the result of testing the classifier for the signal contained in time-bin t only, whereas for the cumulative time-bin analysis, it reflected the result of testing the classifier for the signal over the time-interval $0 - t$, relative to the onset of the test stimulus.

Now, according to a single-process account, there is a unitary signal, reflecting the memory strength composed of the sum of all evidence, which informs the judgement. Thus, if information is integrated over time, we may predict that classifier performance at time t (with the independent time-bin analysis) will be similar to that for time $0 - t$ (from the cumulative time-bin analysis), as Weidemann and Kahana (2019a) found. Alternatively, according to a dual-process account, different signals independently lead to the memory decisions. In that case, we would predict a difference in the classifier performance curves (as function of time) for the independent and the cumulative time-bins. Specifically, if there are independent signals relevant to memory, those should be visible as distinct peaks in the classifier performance curve, when analyzed for the independent time-bins. For the cumulative time-bin analysis, the classifier performance curve will closely follow the more relevant signal for the current time-bin than adding up the evidence from all the previous signals. Figure 1 in Weidemann and Kahana (2019a) presents a schematic illustration of this logic.

Paralleling our results for the LDA weights (see Figure 3.11 for reference), when influences of shorter response-times were not accounted for (or without the vincentization), and for the independent time-bin analysis, classification of old and new trials, perceived-old- and new trials, as well as hits and misses, was chance for the earlier time-bins, then it reached a peak around the 6^{th} time-bin and thereafter dropped back close to chance for the later time-bins (Figure 3.15). The above pattern was true for both LDA and SVM. Thus, very early features of the signals were less likely to contain information relevant to the memory judgments. This trend was also present for the classification of correct rejections and false alarms, but the peak for the independent time-bin analysis was reached even earlier.

For the cumulative time-bin analysis, for all four classifications, both LDA and SVM performances were highly similar to that of the independent time-bins up until they reached the peak, after which performances for the cumulative time-bins approached almost a flat line. Thus, overall, the independent- and cumulative time-bin analyses did not provide

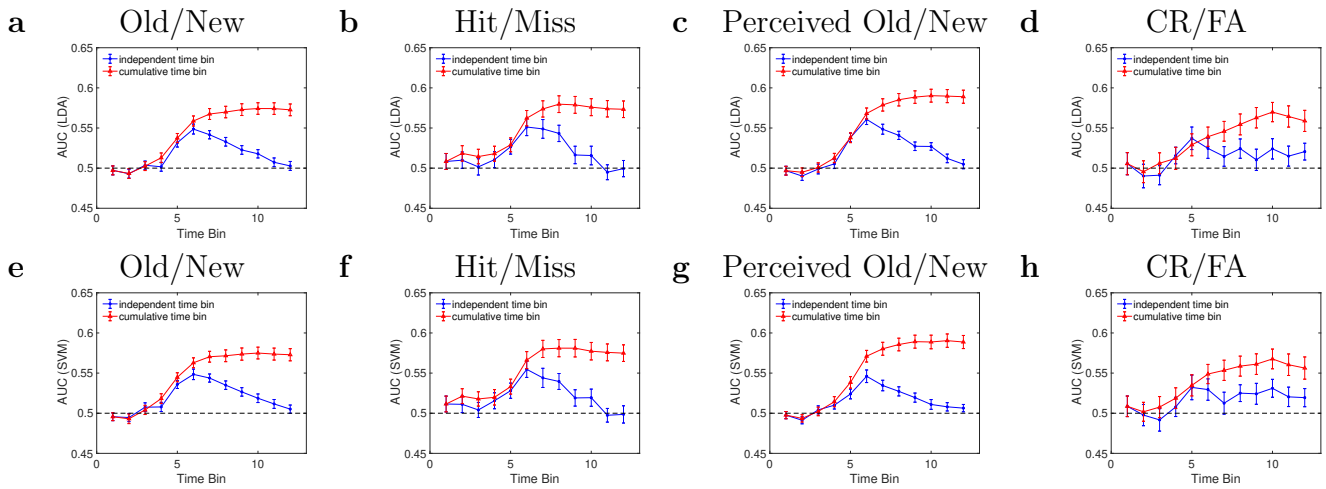

Figure 3.15: Classifications based on independent and cumulative time-bins from 0–1200 ms post stimulus-onset, separately for LDA (upper panels) and SVM (lower panels). For independent time-bin analysis, classifiers were trained and tested with independent 100 ms long time-windows. For the cumulative time-bin analysis, classifiers were trained and tested with sequentially increasing (by 100 ms) time windows. Time features were not corrected for shorter response-times through vincentization. Error bars are standard errors.

evidence for an integrated signal driving the memory judgments.

However, our previous analysis with the LDA weights showed that the peak around the middle time-bin was likely due to average response-times falling right after this interval. Thus, the drop in classifier performance for the second half of the time-bins could be because once the response is made, the participant has no incentive to continue the memory retrieval process, so they may simply be not attending to the probe stimulus at that time. Alternatively, the sigmoid-like performance curve for the cumulative time-bins could be because after reaching a decision, the participant still continues to think about it or perhaps gathers further confidence judgments about the decision. However, since this interpretation is confounded by the response times consideration, we carried out a follow-up analysis with the vincentization method.

With vincentization, the major peak for the independent time-bin analysis went was greatly attenuated (Figure 3.16), supporting that it may have been, at least in part, due to the classifiers tapping into motor-preparatory activity for the response actions.

For the classifications of old and new trials, and the perceived-old- and new trials (Figure 3.16a,c,e,g), classifier performance for the cumulative time-bin analysis (both LDA and SVM) increased with time, almost monotonically. Classifier performance for the independent

time-bin analysis also increased with time, but maintained a difference with classifier performance for the cumulative time-bin, visible after the first few time-bins— the non-overlapping error bars (standard errors) indicate that the difference was significant. Thus, with vincentization, our results were still different from what would be expected for a single-process account.

The difference between classifier performance for the independent- and the cumulative time-bins was even more prominent for the classification of hits and misses (Figure 3.16b,f): for the independent time-bins, it showed a clear early peak, and also a later peak, albeit less clearly; the classifier performance for the cumulative time-bins appeared to be mainly driven by the earlier peak. These results are more in favour of a dual- than single process account.

For the classification of correct rejections and false alarms (Figure 3.16d,h), the classifier performance for the independent time-bins also showed multiple peaks, but the performance curve was less undulating; the performance curve for the cumulative time-bins showed an increasing trend for the earlier time-bins, after which the curve was almost flat. Thus, the early peak seen in performance curve for the independent time-bins may have been important in this case as well.

Taken together, classifier performance curves for the independent- and the cumulative time-bin analyses overlapped with each other only for the first few time-bins, after which the two curves diverged substantially. For the classification of old and new trials, hits and misses, and perceived-old- and new trials, the two performance curves came closer to each other (but did not overlap) again for the very last few time-bins. Also, aside from the initial time-bins, and in comparison to the independent time-bins, classifier performance for the cumulative time-bins was overall higher. This is also suggestive of a dual-process account.

Recall our results without vincentization (Figure 3.15), which showed very similar performance for the independent- and cumulative time-bin analysis, until the major peak was reached. On the other hand, our results with the vincentized signals showed that performance for the cumulative time-bins was different from that for the independent time-bins, and specifically when classifying hits and misses, performance for the cumulative time-bins was mainly driven by an early peak, as seen in the performance for the independent time-bins.

Thus, we wondered if for trials with shorter response times, there was indeed a unitary signal that led up to the decisions whereas for trials with longer response times, this signal was followed by another (independent) memory-relevant signal (as seen for the vincentized

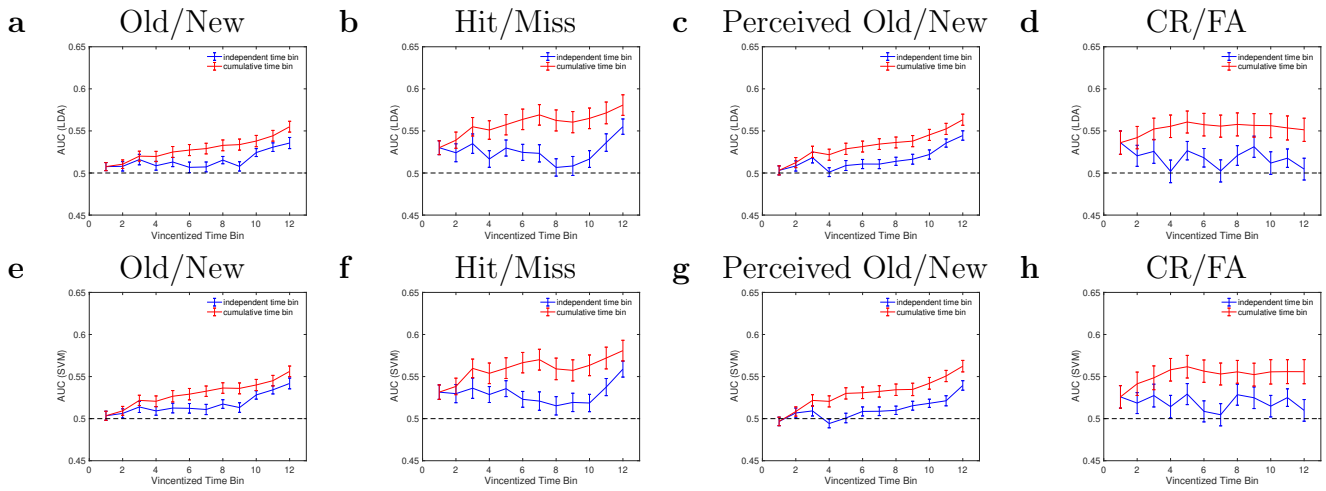

Figure 3.16: Classifications for the vincentized signals, based on independent- and cumulative time-bins from 0–1200 ms post stimulus-onset, separately for LDA (upper panels) and SVM (lower panels). Error bars are standard errors.

signals). We investigated this suggestion with follow-up analyses for the classification of hits and misses, since it showed the most prominent difference between the performance curves for the independent- and cumulative time-bins.

First, we evaluated the classifier performances separately for participants with faster average response times (or those above the median split for average response times) and with slower average response times (or those below the median split for average response times). This showed overall increasing trends in the performance curves, for both independent- and cumulative time-bins, and for the faster participants (Figure 3.17a,c); the error bars overlapped for many of the time-bins and the two performance curves were generally closer together, suggesting a single-process account, at least in part. On the other hand, the performance curves for the slower participants (Figure 3.17b,d) clearly differed for the independent- and cumulative time-bins, and the pattern was more similar to that without separating participants based on their average response times (see Figure 3.16b,f). Together, these results suggest that faster and slower response-times may be linked to the involvement of different memory-relevant signals. For faster participants, there was likely a unitary, integrated signal driving the decisions, which is more in line with a single-process account. On the other hand, for slower participants, an early independent signal was more influential, a late signal was also present but its contribution to the decisions may have been less; this was more in line with a dual-process account.

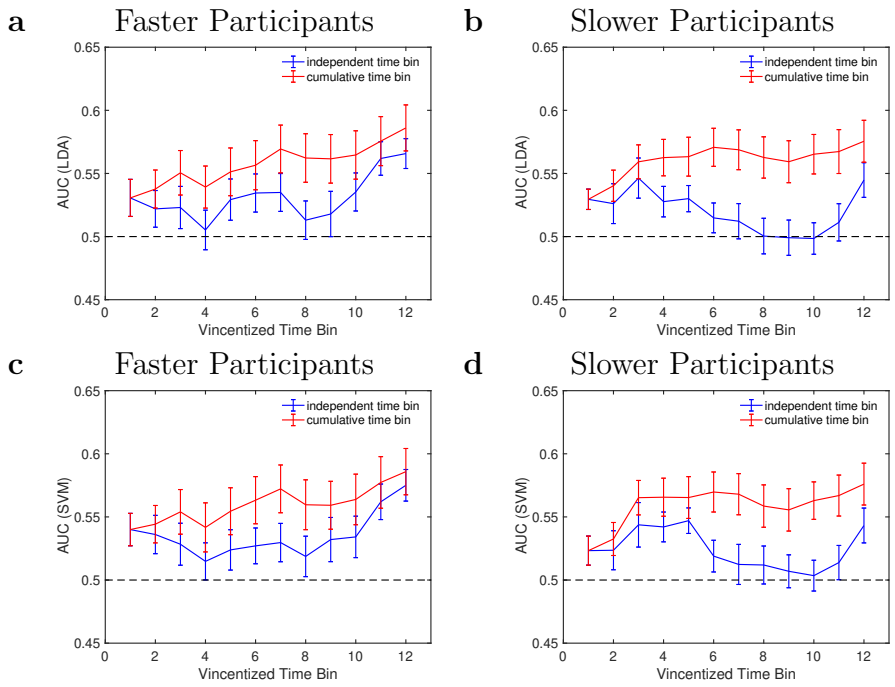

Figure 3.17: Classification of hits and misses based on vincentized signals, separately for independent- and cumulative time-bins, and for LDA (upper panels) and SVM (lower panels). Results are also shown separately for participants with faster- (left panels) and slower average response-times (right panels). Dashed line presents chance (0.5). Error bars are standard errors of the mean.

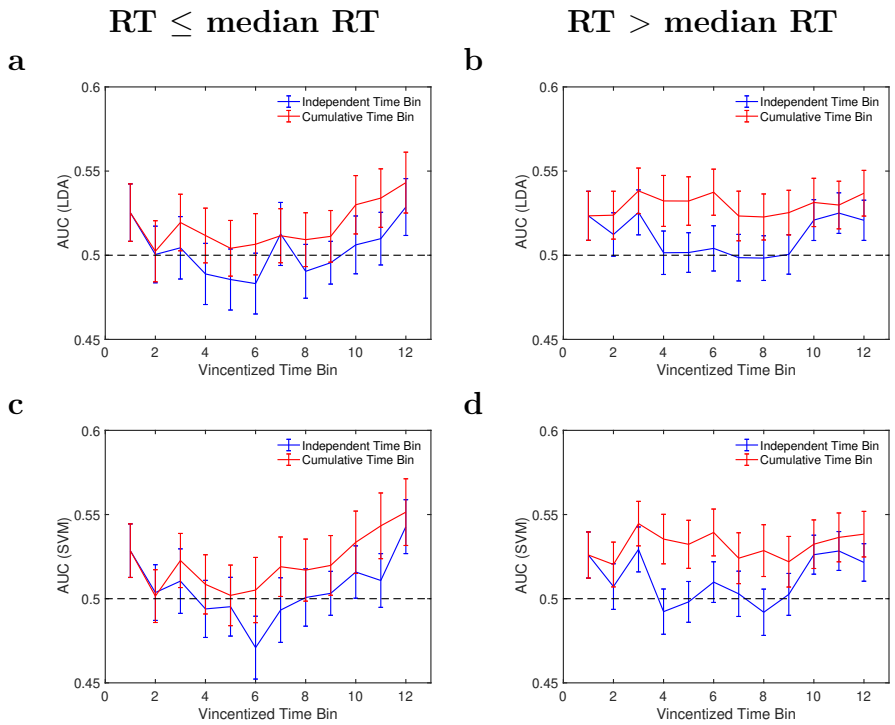

Figure 3.18: Classification of hit/miss trials based on vincentized independent time-bins and cumulative time-bins from 0–1200 ms post stimulus onset, separately for LDA (upper panels) and SVM (lower panels), and separately for trials with response times shorter than the median response time (left panels) and for trials with response times longer than the median response time (right panels). Dashed line presents chance (0.5). Error bars are standard errors of the mean.

Next, instead of splitting the data by participants, we included all participants but split the data by faster versus slower trials. For faster trials (Figure 3.18a,c), indeed the performance curves for the independent- and the cumulative time-bins followed each other more closely (overlapping error bars); both curves roughly increased with time, and reached a peak just prior to making the response, in line with a single-process account. On the other hand, for slower trials (Figure 3.18b,d), the performance curve for the independent time-bins showed an independent earlier as well as a later peak, and similar to the previous analyses, the earlier peak appeared to be relatively more influential. Thus, slower trials, once again showed more support for a dual- than single-process account. Also, one possible explanation for the lesser influence of the later peak relative to the earlier peak could be that the later source of information was highly correlated with the earlier one.

In sum, we found support for both single- and dual-process accounts, respectively for when the participants responded faster and when they responded slowly.

3.3.9 Comparison with the classifiers at study

Finally, we wondered how classifier-based evidence for memory decisions during the test phase compared to that during study, as reported in Chapter 2 (Chakravarty et al., 2020). We only considered the classification of hits and misses, since lures were not present at study. Figure 3.19 presents classifications with all the used univariate (ERP amplitude) and multivariate measures, both at study and at test— namely, 1) LPC and SW amplitude at study, and 2) FN400 and LPP amplitude at test, as well as 3) multivariate activity at study and 4) test, analyzed with LDA and SVM classifiers. There was no significant difference between classification performance with ERP amplitudes at study and at test. Importantly, LDA and SVM at test performed significantly better than 1) LPC and SW amplitudes at study, 2) FN400 and LPP amplitudes at test, as well as 3) LDA and SVM at study ($p < 0.01$ for all cases). Thus, multivariate EEG activity during the test phase predicted memory significantly better than both univariate (ERP amplitude) and multivariate EEG activity at study.

We wondered if study- and test-related activity reflected cognitive processes that independently contributed to memory success at test, in which case, classification of study- and test activity to predict memory outcomes may summate. To test this idea, we ran the classifiers on combined EEG features from the study- and test phases. For each trial, we concatenated the time features from study and test, creating a total of 24 time features per trial (12 from study and 12 from test). Our results showed (Figure 3.19, rightmost bars) that this study+test classifier performed similarly to the multivariate test-activity based classifier, suggesting that study-activity may not contribute to memory variability independent of test-activity. Doubling the number of time-features for the study+test classifier could have produced a smaller effect due to overfitting but equating the number of time-features (12 in total - 6 from study and 6 from test, each averaged over 200 ms time-intervals) did not alter this result. Thus, overall, our results suggested that test activity is more directly predictive of memory in item recognition.

3.4 Discussion

Our main goal for the current study was to investigate if recognition-memory outcomes can be predicted from time-domain features of the EEG signal, present during the test

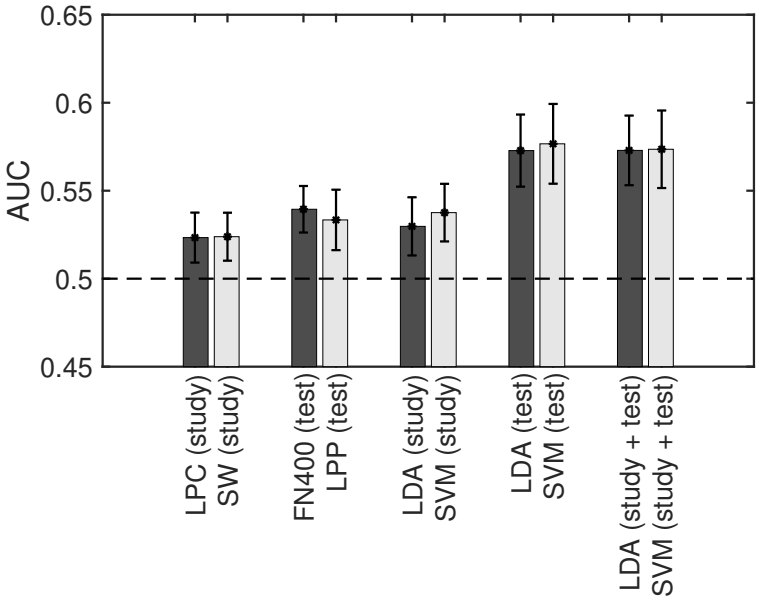

Figure 3.19: Comparison between predictive success with brain activity features from the study face and the test phase, for the classification of hits and misses. Error bars are 95% confidence intervals. Dashed black line refers to chance (0.5).

phase of an item recognition task. We found that 1) univariate ERP (FN400 and LPP) voltage measures, obtained from previous planned-comparisons analysis, predicted memory outcomes significantly better than chance, but the size of prediction was modest overall; 2) multivariate classifiers also achieved modest but significantly better success than the univariate voltage measures, and were associated with individual differences due to better- and worse performing participants. Our classifier approach also led to additional insights about task-related brain activity— 3) a moving-window classifier analysis suggested that the LPP may be epiphenomenal - there could be different activity that reflects the putative recollection (or later-latency) evidence driving the old/new judgment. and 4) test-phase activity predicted memory better than study-phase activity (Chapter 2; Chakravarty et al., 2020). We discuss each of these results in turn.

3.4.1 ERPs at test: FN400 and LPP

Previous trial-averaged ERP effects of FN400 and LPP, such as the old/new effect or the retrieval-success effect, had suggested a monotonic relationship between the amplitudes of FN400 or LPP and memory outcomes. Following up on this suggestion, we found that with the exception of the classification of correct rejections and false alarms, memory outcomes

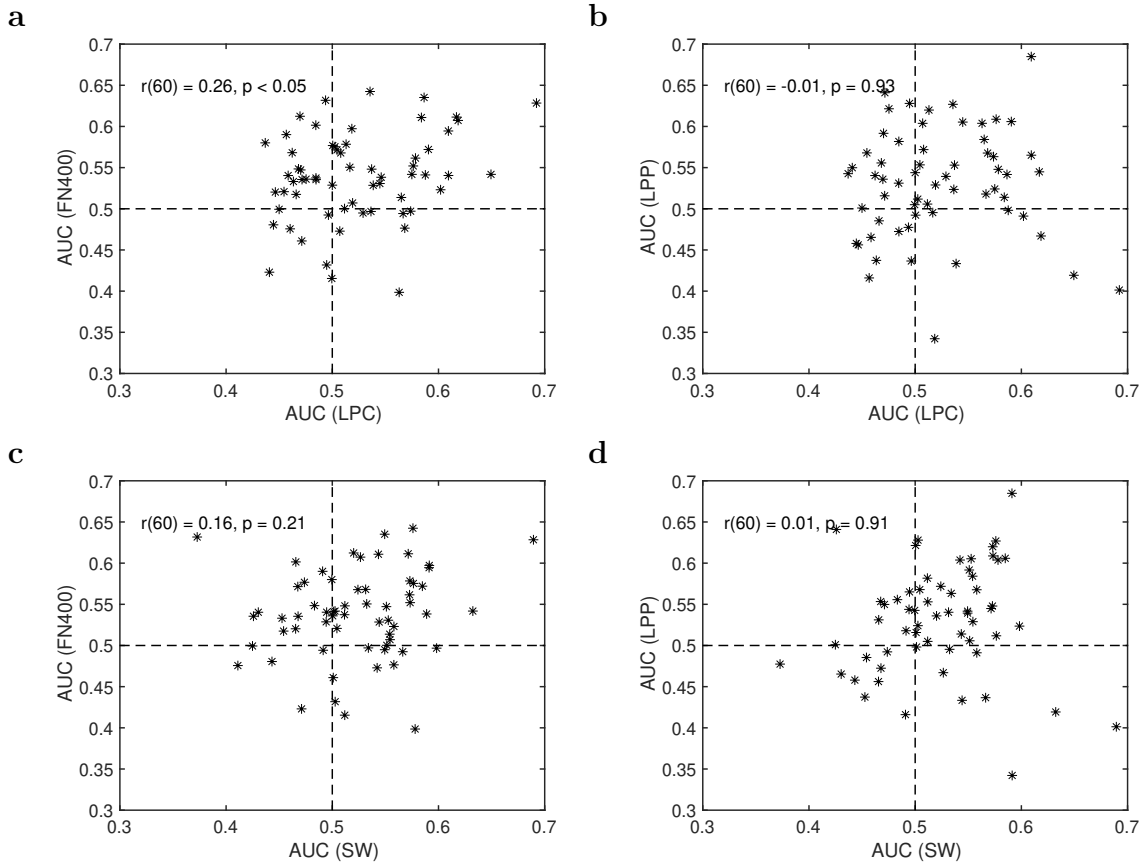

Figure 3.20: Correlation between predictions (AUCs) obtained with ERPs at study (LPC and SW) and ERPs at test (FN400 and LPP), for the classification of hits and misses. Dashed black lines refer to chance (0.5).

could be predicted modestly, at the level of individual trials, based on the amplitudes of the two ERPs. The failure to predict correct rejections and false alarms with FN400 or LPP amplitude suggested that these two signals may not index false alarms in a similar way as the targets than the lures, as has been suggested previously by some researchers, specifically for the FN400 (Finnigan et al., 2002; Wolk et al., 2006). Also, for the other three classifications considered here, for which both FN400 and LPP amplitudes led to successful predictions, we noted the following important points.

FN400 and LPP may be functionally different signals Predictions based on FN400 and LPP did not correlate with each other, for any of the classification problems. This paralleled the results from traditional ERP analysis of the FN400 and LPP in Chen et al. (2014), who found that at the trial-averaged level, FN400 and LPP amplitudes did not correlate with each other, either when considering the difference wave for the old/new contrast or the hit/miss contrast.

FN400 may be more relevant to memory success FN400 amplitudes classified hits and misses significantly better than its classification of old and new trials, or perceived-old- and new trials. Also, participants' d' correlated with predictions for hits and misses with the FN400 amplitude. In parallel, Chen et al. (2014) also found correlations between d' and the trial-averaged amplitude of the FN400, when considering the difference wave for the hit/miss contrast. Further, we found that predictions based on FN400 amplitudes were not significantly altered when trials with faster response times, with respect to the latency of the FN400, were excluded. Lastly, as we discuss later in this section, even for the slow responses, the multivariate classifier analysis showed that the early source of evidence was deemed more influential for the memory decisions. All of these results suggest that for this task, FN400 may have been more relevant to memory success.

LPP may be epiphenomenal for old/new item recognition Classification of old and new trials, or perceived-old- and new trials, were significantly better with the LPP than the FN400 amplitudes. However, when faster responses, relative to the latency of LPP, were excluded, predictions with LPP amplitudes became worse, except for the classification of correct rejections and false alarms, which became significant. This could be due to reduced

power, for many trials were excluded when considering faster responses relative to the latency of LPP. Alternatively, for faster responses, LPP-indexed recollection may have been more likely present, but it was epiphenomenal, for the decision had already been made, and was likely driven by familiarity. Further, as we discuss below, for slower responses, the multivariate classifier analysis showed that although a late source of evidence was present, it did not influence the decision as much as the earlier source of evidence. Taken together, LPP-indexed true recollection process may have been rare in this task, and when present, it may have added more to the meta-judgments about the decisions rather than exerting a great influence on the decisions themselves.

In sum, our analyses with FN400 and LPP amplitudes achieved overall significant success in predicting memory outcomes but maintained the suggestion that contributions of these two ERP features in memory may be small, and also unrelated to each other.

3.4.2 Multivariate pattern analysis of brain activity at test

Multivariate pattern analysis of test-phase activity achieved significant success for all four classification problems that were of interest for this study, though the size of prediction was still modest. Our size of predictions were similar to a recent study by Noh et al. (2018), though the goals of the two studies were different. The average classifier accuracy for Noh et al. (2018) was 61%, and since they had balanced the training classes, classifier accuracy could be compared with the classifier AUCs we reported here; so for example, the average AUC, for our classification of targets and lures was 0.58, which is comparable to their average classifier accuracy.

Memory-relevant signals beyond FN400 and LPP The classifiers predicted significantly better than FN400 or LPP amplitudes alone. This was possible, for the classifiers were able to find combinations of different features. It may also add further support to the above suggestion that FN400 and LPP may not contribute to common memory variability. Moreover, the classifiers also succeeded in predicting correct rejections and false alarms, which was not successful with the FN400 or LPP amplitudes, at least when all response times were considered. Thus, there may have been other relevant signals, beyond FN400 or LPP, that drive the difference between processing of false alarms and correct rejections. Supporting this suggestion, analysis of the LDA weights for the classification of correct re-

jections and false alarms showed greater influence of the bilateral temporal scalp regions for the 500–600 ms time-bin post stimulus-onset, while all three other classifications (old and new; perceived-old- and new; hits and misses) relied on an LPP-like signal, indexed by higher LDA weights in the posterior scalp.

Interestingly, classifier results from Noh et al. (2018), who directly used FN400 and LPP amplitudes (across different electrodes) as the features for the multivariate classification analysis, suggested that false alarms may be processed more similar to targets than lures. Their suggestion was based on the similarity between classifier scores for false alarms and hits, than for false alarms and correct rejections. However, the corresponding classifier was trained and tested for the difference between hits with incorrect source information and correct rejections. Thus, false alarms may better scale with a subset of the hit responses, for which additional details cannot be retrieved correctly.

Individual differences due to performance As mentioned above, the classifiers achieved significant but modest success in predicting different memory outcomes. However, classifier performance showed positive trends with participants' d' and the correlations were significant in some cases, for example, the classification of old and new trials. Thus, despite the overall modest size of prediction, classifier performance for better-performing participants was meaningfully large. A possible interpretation for the positive trends between classifier performance and participants' d' could be that for the better-performing participants, who may have been well-motivated to do the task or did not find the task particularly difficult, the corresponding brain activity had more task-relevance (or "task-resolution"). This higher task-resolution of brain activity for the better-performing participants may have been beneficial for the classifiers, which could better pickup on the signal changes, and thus, make better predictions for the behavioural outcomes. This suggestion is further supported in Chapter 2 (Chakravarty et al., 2020) and also by Arora et al. (2018), who also found correlations between classifier- and participants' performance.

Influence of motor-preparatory activity Because classifiers are data-driven techniques, they can be influenced by circumstantial evidence. In the case of test-phase activity, motor-preparatory activity for making the responses, may provide circumstantial evidence for the classifier analysis. For example, behavioural analysis of the current data (see Chen et al.,

2014) showed that hits were significantly faster than misses. Thus, the classifier may be able to pick up this information and use it to classify hits and misses. Shorter response times, relative to the chosen epoch length for the trials, are more likely to include motor-preparatory activity, and analysis of the average response times for the current data showed that indeed many responses fell short of the chosen 1200 ms time-window of analysis, and thus, could have included motor-preparatory activity. However, classifier analysis with the vincentized signal, which did not include motor-preparatory activity, produced similar predictions as that without the vincentization. Thus, although likely present, motor-preparatory activity did not substantially influence the classifiers. Instead, meaningful predictions were made from brain activity prior to when the response was made.

Single- and dual-process accounts of recognition memory As mentioned before, single- and dual-process accounts of recognition memory suggest that memory judgments are either driven by unitary, integrated source of evidence, or by two (or more) independent sources of evidence, respectively (Dunn, 2008; Wixted & Stretch, 2004; Wixted & Mickes, 2010; Yonelinas, 1997, 2002). The traditional ERP effects of FN400 and LPP have added to this debate because the two ERPs are commonly modulated by different experimental factors and thus, they may support different sources of information. However, those different sources may still be integrated into a unitary evidence for recognition judgments. Importantly, with the trial-averaged ERPs, we cannot examine the amount of neural evidence behind memory decisions, which would be more useful to understand the plausibility of single- and dual-process accounts.

Using classifiers to trace the amount of neural evidence over time, Weidemann and Kahana (2019a) found a single process account to be more plausible. However, we found different results. Considering the vincentized signal, classifier performance, for the cumulative- and independent time-bins overlapped with each other for the initial few time-bins only, after which classifier performance for the cumulative time-bins was greater. Classifier performance for the independent time-bins showed an early and a late source of evidence, which was seen more clearly for the classification of hits and misses. The early source was more influential, for the classifier performance for cumulative time-bins was mainly driven by this early source. Thus, in contrast to Weidemann and Kahana (2019a), our results were more in line with a dual- than single-process account.

Consideration of fast and slow responses further informed this result. Different from the trend we found with the vincentized signal, when vincentization was not included, the classifier performance for the cumulative and independent time-bins completely overlapped with each other until the single peak for independent time-bins was reached, which was likely produced because decisions were typically reached around this time. Based on this finding, and the above-mentioned trend for the vincentized signal, we wondered if fast responses were made based on a single process, whereas some slower responses might be driven by the two putative, distinguishable sources of evidence. Supporting this suggestion, classifier performance for both the cumulative- and independent time-bins increased with time and followed each other more closely (with overlapping error bars) for faster than slower participants, or for faster than slower responses. For slower participants/responses, the above-mentioned trends with the vincentized signal reappeared. Thus, slower responses were supported by two sources of evidence, though the earlier source was more important. In other words, we found that both single- and dual-process accounts could be relevant, but for fast and slow responses, respectively. We also found that the influence of the later source of evidence, for the slower responses, may less than that of the earlier source. This result may be reconciled with Dunn (2008) who used state-trace analysis to suggest that there is not enough evidence that the remember/know task recruits qualitatively different sources of memory-relevant information. Thus, it is possible that recollection-based judgments are too rare to produce non-monotonicities in state-traces.

This raises the question, why do our results diverge from Weidemann and Kahana (2019a)? Weidemann and Kahana (2019a) classified targets and lures. In our case, the difference in classifier performance for independent- and cumulative time-bins was much smaller when classifying targets and lures, than when classifying hits and misses (see Figure 3.16a,e). When classifying targets and lures, classifier performance for both independent- and cumulative time-bins, roughly increased with time, and at the same rate. Thus, considering the classification of targets and lures only, our results may not have been very different from Weidemann and Kahana (2019a). Second, Weidemann and Kahana (2019a) used spectral classifiers. Spectral estimates are computed over time-windows centered around a particular time. This makes their time-bins effectively longer, and with more overlap in the moving window analysis, potentially making it less different than the cumulative window than they would have liked. Another difference between the two studies is that as the number of fea-

tures sequentially increased for the analysis with the cumulative time-bins, Weidemann and Kahana (2019a) changed the amount of regularization in the model (SVM) relative to the increasing number of features, whereas we kept it constant, because amount of regularization relative to number of features do not follow a deterministic rule. Thus, our classifiers may have under-performed for the independent time-bins. However, in general, we have found our results to be robust to a fixed amount of regularization or individually tuned amount of regularization (e.g., see Figure 7 in Chapter 2; Chakravarty et al., 2020), possibly because our predictions were overall small in size. For example, for classification between targets and lures, the average AUC in Weidemann and Kahana (2019a) was 0.71, whereas the average AUC for our study was 0.58. It is possible that the spectral measures of EEG used as features in Weidemann and Kahana (2019a) offered a better SNR than the time-domain voltage measures used in the current study.

Others have also taken related approaches to evaluate single- and dual-process accounts. For example, with multiple regression analysis, Ratcliff, Sederberg, Smith, and Childers (2016) suggested that old/new recognition judgments may be driven by LPP only. Their study did not directly predict old and new trials with the EEG activity, but instead it predicted the trial-to-trial drift rates (from a fitted diffusion model). However, the use of drift rates could be useful as these measures integrate both accuracy and response time information for each trial. Another study by van Vugt, Brandt, and Schulze-Bonhage (2017), who used the intra-cranial EEG signal, filtered in the 4–9 Hz theta band and binned into 50 ms long time-bins, found that performance of a logistic regression classifier, for the classification of old and new responses, increased with time, and interestingly, it kept increasing even after reaching the response. It is possible that the participants were still gathering confidence judgments about the recently made decision even though confidence ratings were not asked for. Alternatively, since they did not correct for motor-preparatory activity, which was likely present in the signal, the increase in classifier performance may only be due to the classifiers picking up on circumstantial evidence. Notably, the performance of the classifiers in van Vugt et al. (2017) was more comparable to our study; their median AUC was close to 0.63.

3.4.3 Comparison between study- and test-phase activity

Overall, our results provide a detailed account of classifying the time-domain EEG signal during the test phase of an item recognition task. Consolidating with Chapter 2 (Chakravarty

et al., 2020), we found that classifying hits and misses using the two ERP features at study, namely LPC and SW, was not significantly better than the same using the features of the two ERPs at test, FN400 and LPP (Figure 3.19). On the other hand, multivariate features at test produced predictions that were significantly better than the same for ERPs at test (FN400 and LPP) or at study (LPC and SW) or even the multivariate features at study. Taken together, this suggests that brain activity at test is more relevant to the recognition judgments. This is possible, because in old/new recognition judgments, hits versus misses will depend on placement of the response criterion, and response criterion will depend on the relative strength distributions for targets versus lures; thus, it stands to reason that test activity will have a more complete picture of what drives the decision.

Interestingly, we found that across participants, LPC amplitude based predictions at study were correlated with FN400 amplitude based predictions at test (Figure 3.20a), suggesting that FN400 and LPC may share common variance in explaining memory outcomes. Previously, the ERP analysis of the current data had also found correlations between FN400 and LPC amplitudes for the retrieval success effect (see Chen et al., 2014). Thus, one possibility, as suggested by Chen et al. (2014), is that memory variability due to cognitive processes at study are retrieved by those at test. In addition, test-related processes also contribute to the memory variance on their own. Together, this will produce better memory predictions for test-related activity than study-related activity. Alternatively, despite a shared variance, study-related processes may also contribute to memory variance than is not shared by test-related processes.

Based on the above ideas, we tested the multivariate classifiers for the classification of hits and misses, with features from both study- and test-phases, concatenated together. This *study+test* classifier predicted similarly to the test-activity classifier. Thus, consistent with Chen and colleagues' suggestion, test-phase activity included the variability in memory due to study-phase activity as well as had its own contribution to the variability, e.g., the variability due to the lures, which were never part of the study, but were an important determinant of memory performance.

3.4.4 Other considerations

To improve the classifier performance, a few other considerations also seem relevant.

Spectral- over time-domain features of EEG The time-domain features of EEG that were used here for the classification are likely influenced by variability in the latency of the signals from one trial to another, in comparison to the spectral features, which use larger time-windows to compute the power estimates and thus, are less influenced by trial-to-trial variability in latency. Thus, using spectral-features for the classification analysis could be a future direction for us, as has also been followed by some other studies (Ezzyat et al., 2017; Noh et al., 2014; van Vugt et al., 2017; Weidemann et al., 2019; Weidemann & Kahana, 2019a). However, the temporal resolution of the signal is higher with the time- than the spectral features, which could be useful in situations, for example, when estimating occurrence of motor-preparatory activity in the signal more accurately. Also, both time- and spectral-classifiers could be leveraged to find overall better predictions (e.g., Noh et al., 2014).

Choice of classifiers The two classifiers used here, LDA and SVM, were suitable for discriminating between classes that are linearly separable, and were chosen to respect our plan for progressing systematically to avoid overfitting at the level of the classifier algorithm (Skocik et al., 2016). Although linear classifiers like SVM has seen frequent success and is a common choice in cognitive neuroscientific research, it is possible that memory-related brain activity patterns are better explained with the help of non-linear models, which also account for interactions between the features (e.g., Sun et al., 2016). Thus, another future direction could be to use non-linear classifiers (Arora et al., 2018; Sun et al., 2016). However, non-linear models lack interpretability and thus, we may not be able draw inferences about the characteristics of behaviourally-relevant brain activity, as identified by the classifiers. Further, depending on number of features, non-linear models require a sufficiently large number of trials to learn their main and interaction effects, whereas memory experiments include at most a few hundreds of trials.

Class imbalance For classifying hits and misses, as well as for classifying correct rejections and false alarms, imbalance between the classes can be a potential challenge for the analysis. With the SMOTE oversampling method used in this study as well as in Chapter 2 (Chakravarty et al., 2020), across participants, we did not find significantly different classification results for with and without balancing the trials. However, other ways to address the

class-imbalance could produce different results. Also, oversampling is limited to the training sets only, and thus, test sets are still imbalanced. We ended up excluding cases where one or more test folds contained examples from the major class only, and thus, ROC and AUC of the ROC could not be computed.

Correcting for motor-preparatory activity with vincentization The vincentization method for the time-domain signal, depending on the response time, could have averaged the ERP waveforms over very small or very long time-intervals, whereas the typical duration of ERP signals are a couple of hundred milliseconds only. Thus, vincentization may have also influenced our results and a different approach to account for motor-preparatory activity could produce different results. Thus, our results provide benchmarks of predictive accuracy for the test-related EEG signals, which could be compared against in follow-up investigations exploring methodological improvements to our analysis.

Other predictors of memory Different from the possible methodological improvements, an alternative reason for the modest predictions could be that there exists other important predictors of memory which are absent from the investigations presented here. To our benefit, behavioural research over the last few decades has identified many factors that influence memory (Humphreys et al., 1989; Kahana, 2012; Lewis, 1979; Surprenant & Neath, 2013). Many of these factors that determine memory success could lie outside of the study or the test phase. For example, researchers have also found activity during the delay between study and test (e.g., Polyn et al., 2005) as well as the pre-stimulus activity (e.g., Park & Rugg, 2010) to be relevant to memory outcomes. Thus, investigations with the classifiers that include these other predictors in an incremental fashion could offer valuable insights into memory processes.

Applications for the future The predictive models offer interesting applications for memory training. For example, classifier-identified patterns of brain activity during the study phase could be used to reinforce activity that leads to successful remembering, with a neurofeedback setup, helping people self-regulate into states more conducive of memory success. A recent study by Ezzyat et al. (2017) found that classifier-contingent neuro-stimulation at study could lead to better memory at test, which may partly support the potential success

of the above-mentioned neurofeedback setup. Classifier-identified patterns of brain activity from the test phase could also be used in applications such as brain-computer interfaces, to communicate the participants decisions, when direct behavioural responses are not possible to obtain.

3.4.5 Conclusion

Building on previous work showing that brain activity, both during the study- as well as the test phase index memory outcomes in item recognition, we show that test-related activity may be a better predictor than study-related activity and that classifiers applied on test-related activity can also offer insights into cognitive processes that guide memory.

Chapter 4

An event-related potential analysis of trial-and-error learning

Abstract

Convergent evidence suggests that the feedback-related negativity (FRN) indexes reward-prediction error (RPE; e.g., Holroyd & Coles, 2002). FRN studies have typically used learning problems like the gambling task— given an extremely small number of stimuli, one attempts to find an efficient response strategy to maximize total rewards, because outcomes are probabilistic. We tested whether the RPE role of the FRN generalizes when reward prediction is not the main goal but rather, guides participants in a challenging stimulus-specific learning task. Participants had learned whether or not to choose each of 48 words depending on each word's inferred value (high or low). Over 16 repetitions, most participants learned the response rules with very high accuracy. Next, we introduced a surprise reversal that randomly toggled half of the words, to induce strong expectation-violations. A signal resembling the latency and polarity of the FRN was elicited and was relevant for feedback-mediated learning, but this signal was more frontal in its scalp distribution of voltage than a mid-frontal negativity, which is thought to be a characteristic of the FRN. Moreover, contradicting our hypothesis, amplitude of this FRN-like signal was not significantly greater for switched than non-switched words. Follow-up analyses identified RPE-like response properties of FRN-like activity, but this was influenced by task variables that were not designed to manipulate RPE, such as word value. The FRN-like signal was associated with response adjustments related to updating learning, but this also depended on response strategies spontaneously adopted by the participants during the surprise reversal. Taken together, even when predicting reward is not the participant's primary goal but it guides learning, the FRN, or a signal that partly resembles the FRN, may be present. However, the mapping of this FRN-like signal onto RPE may be intimately modulated by the fine structure of the task, deviating from the abstract, task-independent formulation of RPE of reinforcement learning.

4.1 Introduction

Ample research suggests that rewards play important roles in learning and decision making. Reinforcement learning (RL) theories model learning that takes place through trial-and-error and with the ultimate goal of maximizing the total reward (Sutton & Barto, 1998). RL is an exciting growing front of research for experimental investigations as well as for development of machine intelligence algorithms. Moreover, animal models (e.g., Schultz et al., 1997) and human neuroimaging (e.g., Pessiglione et al., 2006) suggest biological plausibility of the RL framework, leading to an overwhelming number of studies published in the last two decades itself, examining reward-driven behaviours and their neural correlates. The majority of the behavioural paradigms used to study such questions are inspired by the earlier animal studies (e.g., Lau & Glimcher, 2005).

However, while research in the field of computational RL has explored and modelled a wide range of learning situations, cognitive neuroscientific studies have focused on rather simpler tasks to study human reinforcement learning. Accordingly, it is important to ask if the previously found RL-relevant brain-activity signals retain their functions beyond the common experimental paradigms. Here, we looked into an EEG signal, namely the feedback-related negativity (FRN), which is commonly thought to reflect the instantiaion of the RL framework. However, FRN research has mainly focused on the two-armed bandit problem (explained in detail below). In contrast, respecting the broader domain of learning situations in RL and also borrowing inspirations from the verbal memory literature, we used an RL problem where the responses are relatively simple to learn but the participant is challenged with remembering the responses mappings of a large set of stimuli (words). To better set the current study in context, first, we briefly review the basic RL principles and evidence for a biological RL framework, including a background for the FRN. This is followed by a comparison between other commonly used FRN paradigms and the current task. Finally, we note FRN-relevant hypothesis for this study.

In RL, one learns to respond optimally for a 'state' in multiple steps or 'actions'. A 'reward function' is evaluated at each of these steps. The state, action and reward are the basic components of RL (Schultz, 2015). In experiments, state can sometime refer to properties of the stimuli, or it can also be an internal variable, not directly observable. Actions are response/choices; and the reward function evaluates the quality of the action with

respect to the stimuli. Notably, state, action or even reward variables can be quite intricate (e.g., in real-world situations), increasing the complexity of the RL problem. However, the basic RL framework makes simpler assumptions for these. When accounting for more complex RL problems, researchers have often stressed on the importance of neuro-cognitive functions such as attention, memory and executive functions (e.g., see Gershman & Daw, 2017; Rmus, McDougle, & Collins, 2021). As mentioned above, the goal of RL is to maximize the reward in the longer term. The reward function in RL is characterized by reward-prediction error (RPE)— the difference between the prediction about the outcome and the actual outcome. Predictions arise from past experiences and refer to the internal expectations about the outcomes. RPE can guide future actions, these computations are trial specific, and are used to sequentially update the predicted reward value for the next trial. Accordingly, the action for the future trial is also updated. Then, across trials, depending on the rate of learning, the gap between the predicted and the actual reward value approaches the minimum. In other words, when the outcome is worse than expected, response is adjusted in the future trial; when the outcome is better that expected, response remains the same (is learned) for the future trial. As learning takes place, predictions become stronger and closely resemble the actual outcomes.

Across many studies, researchers have found support for RL-relevant computations in the brain. An early influential finding is the activity of the dopamine neurons in the midbrain, which show phasic responses to RPE. Animal studies showed that when the outcome is better than predicted (or in the case of a positive RPE), there is a phasic increase in the firing rates of the dopamine neurons whereas when the outcome is worse than predicted (or in the case of negative RPE) there is a phasic decrease (from spontaneous rate) in the firing rates of dopamine neurons. When the outcome is similar to that predicted (zero RPE) no such changes in dopamine activity was observed (Schultz et al., 1997; Tobler et al., 2005). Thus, the dopaminergic responses in animal midbrain may be bidirectional, and reflect (signed) RPE (Hart, Rutledge, Glimcher, & Phillips, 2014). Analogous human brain regions may be involved in reward predictions as well. For example, using neuroimaging techniques, various brain structures in the human midbrain have been linked to RPE, specifically the BOLD activity levels code for RPE in the the ventral striatum (Pessiglione et al., 2006) and ventral tegmental area (VTA; D'Ardenne et al., 2008). In addition, neurochemically, human dopaminergic response may also reflect RPE, since administration of drugs that change

striatal dopamine activity accompanied the changes in the RL action variable (Pessiglione et al., 2006). Therefore, a dominant idea is that the mesolimbic dopamine pathway facilitates learning by inducing synaptic plasticity in the striatal (midbrain) regions.

The dopamine neurons of the midbrain also project on to higher order cortical areas. It has been suggested that the communication between dopamine-indexed RPE with cortical areas mediate control of action/response choices. Specifically, the motor areas near the cingulate cortex are thought to play active roles in the inhibition of responses (Kerns et al., 2004). However, recent evidence has been suggesting that the role of higher order cortical areas, specifically the prefrontal cortex (PFC) is much more subtle (Seamans & Yang, 2004) than simply relaying information from the midbrain dopamine systems. The PFC may have a more direct role in reward-driven learning, supporting mechanisms that resemble dopamine-based computations but are less likely to be entirely driven by the midbrain dopamine activity (Rushworth & Behrens, 2008), particularly when the learning variables or the tasks are more complex (Barraclough, Conroy, & Lee, 2004).

With scalp-recorded EEG, researchers have identified a signal, elicited during processing of the outcome and is likely generated from the anterior cingulate cortex (ACC) (Hauser et al., 2014; Gehring & Willoughby, 2002; Van Veen et al., 2004; Miltner et al., 1997), a site that is thought to index RPE (Rolls et al., 2008). This signal, known as the feedback-related negativity (FRN), is a negative-going deflection with a latency of 200–350 ms post onset of the feedback stimuli (Miltner et al., 1997; Nieuwenhuis et al., 2004). The scalp-distribution of voltage for the FRN has a characteristic mid-frontal negativity. Across multiple studies, it has been observed that the FRN has a greater deflection following negative than positive feedback outcomes (Bellebaum & Daum, 2008; Holroyd et al., 2004; Pfabigan et al., 2015). Other relevant connections between the FRN amplitude and properties of the feedback signal have also been reported, such as sensitivity to reward/monetary feedback (Yeung & Sanfey, 2004), win or loss outcomes (Gehring & Willoughby, 2002; Holroyd & Coles, 2002; Yeung, Botvinick, & Cohen, 2004; Frank, Woroch, & Curran, 2005), small versus large reward outcomes or reward versus punishment outcomes (Yeung, Botvinick, & Cohen, 2004; Wu & Zhou, 2009; but see Goyer et al., 2008 for a contrasting finding), reward probability (Nieuwenhuis et al., 2002; Holroyd & Coles, 2002; Goyer et al., 2008; Walsh & Anderson, 2011a; but see Oliveira, McDonald, & Goodman, 2007; Chase et al., 2011; Yu, Zhou, & Zhou, 2011; Wu & Zhou, 2009 who refute this) etc.

Taken together, these results suggest a central place for this signal within the RL framework. The FRN-RPE account (Holroyd & Coles, 2002) predicts a larger (more negative) FRN when the outcome is worse than predicted in comparison to the same for when the outcome is similar to that predicted (for a meta-analysis, see Sambrook & Goslin, 2015). This account is motivated by the potential contribution of the ACC in generating this signal.[1] Consider that if the dopamine neurons of the mesocortical pathway produce an inhibitory effect on the neuronal population in and surrounding the ACC region, then a phasic increase of the dopamine activity would inhibit the ACC neurons, which in turn will produce more positive FRN. In contrast, a phasic decrease in dopamine activity would disinhibit the ACC neurons, which in turn will produce more negative FRN. Note that some recent studies suggest that underlying the FRN, there may actually be two signals: the N2 deflection in response to unexpected outcomes and a positive going deflection (also known as the correct-related positivity) which relates to the processing of positive outcomes (Holroyd, Pakzad-Vaezi, & Krigolson, 2008). Thus, when the outcome is negative, the superposition of these two signals with different polarity will produce a more negative FRN than when the outcome is positive (for a review, see, Proudfit, 2015).

Different from research in the field of computational reinforcement learning, cognitive neuroscientific research has focused on a specific class of experimental paradigms to understand the functional significance of the FRN. For example, a classic RL paradigm that is commonly used to test FRN effects is the two-armed bandit task (Gehring & Willoughby, 2002; Hajcak et al., 2007; Goyer et al., 2008). Motivated from the situation of a gambler operating on two different slot machines, here the participant has to bet on two different stimuli that give out rewards with a specific probability (e.g., 80% versus 20%). Thus, on each trial, there is always some uncertainty associated with the reward, even after the participant has learned which stimulus has a greater payout probability. In other words, even after learning that stimulus A gives out the reward about 80% of the time and that for stimulus B is 20%, the participant still needs to guess on each trial whether the reward might be under A (exploit) or B (explore), in order to maximize the long-term rewards. Computational solutions to this problem, such as the ϵ–greedy solution suggests that choosing to explore every once in a while is better than never choosing to explore. Humans are also likely to use

[1] Though some have suggested the source of the FRN to be in the basal ganglia (Martin, Potts, Burton, & Montague, 2009).

different strategies while performing this task. However, arguably, the task design is more detailed for the response variable than for the stimulus. In cognitive neuroscientific research, the 'gambling tasks' have typically used a small number of stimuli (e.g., two to four options) and as such there is little challenge to remember them, more so when stimuli are paired together and the better option may be remembered at the expense of the other.

Another popular task specific to FRN research is the time estimation task (Miltner et al., 1997; Nieuwenhuis, Slagter, Von Geusau, Heslenfeld, & Holroyd, 2005; Gehring, Goss, Coles, Meyer, & Donchin, 1993). Here, the participant is asked to estimate whether a specific time interval (e.g., 1 s) has passed and press a button to indicate it. Feedback informs about the accuracy of the response and also the margin of error. Over repetitions, the participant learns to minimize this margin. Usually a response criterion is set. For example, all responses that fall within a small, fixed margin of error are considered accurate; all responses outside of this margin are considered inaccurate. However, this criterion could become stricter when the responses become mostly accurate. Thus, the task is set up in a way so as to make the number of accurate and inaccurate responses comparable. Similar to the gambling task, here too the trial-by-trial rewards are associated with some uncertainty and the shifting response-criterion helps maintain it even when the participant has learned to respond well. Importantly, here learning reflects a conditioned motor response, which is shaped by the feedback.

The FRN has also been suggested to guide RPE-driven response adjustments (for a review, see Luft, 2014), as does the RPE in an RL problem. For example, van der Helden, Boksem, and Blom (2010) found that in motor sequence learning, negative feedback to incorrect responses that were subsequently corrected showed a larger FRN than those that were not followed by correct responses. Gambling tasks also show larger FRN for the high-rewarding stimulus earlier than later in the task, suggesting a smaller RPE due to learning which in turn could suggest that the participants come to expect a reward more strongly for the high-rewarding stimulus. Also, for the low-rewarding stimulus, the effect is opposite, suggesting that participants come to expect the reward less strongly for the low-rewarding stimulus (Cohen, Elger, & Ranganath, 2007). Interestingly, recent behavioural evidence also suggests that prediction errors may not be exclusive to RL and could also support learning even with single exposure to the items (e.g., declarative learning; see Greve, Cooper, Kaula, Anderson, & Henson, 2017).

In an effort to isolate RPE and keep experimental design simple, cognitive neuroscience efforts to identify neural correlates of RPE have only scratched the surface of the range of learning problems investigated in the field of RL. A clear limitation of the tasks considered thus far is that the learning challenge is either absent or nearly trivial. Accordingly, here we consider an RL task where the number of stimuli to be learned in the stimulus–response space is substantially greater ($N = 48$). The rewards are deterministic, and accordingly, the policy that guides the choice given a stimulus is also not stochastic, as in the case of the gambling or the time estimation task. Instead here the policy is finding the map $f : S_i \rightarrow R_i$, where $R \in \{choose, not\ choose\}$. However, with a large number of stimuli, memory for a particular stimulus–response pair $\{S_i, R_i\}$ is likely to be interfered with the memory for another pair $\{S_j, R_j\}$, and the interference can be reduced with repetitions of the pairs. This also prevents the RPE function from converging to zero very quickly and produces learning curves (pooled across all the stimuli) that are comparable to more common RL paradigms. Also note that unlike the gambling paradigm, here the stimuli are not paired with each other and thus learning about one stimulus does not help with the choice for another. One way this situation could be different from that of a gambling task is that in order to do better, here the participant needs to remember more detailed information about the stimuli, which may recruit higher order cortical areas (such as those for declarative memory) more than when there are only very few options to make choices for. In computational RL research, the calculation of RPE follows the same mathematical function irrespective of the nature of the task or the number of the stimuli. In cognitive neuroscientific research the mapping of RPE onto the FRN has held up fairly well thus far, and so it is possible that there is an all-purpose RPE-estimator in the brain, corresponding to the FRN. However, given the limited scope of tasks examined, it is also quite plausible that RPE might be "situated" or computed differently, even by different brain regions, depending on the task. It is also possible that although RPE might be part of the inherent calculations performed by the brain in more complex tasks, there might not be a single brain-activity signal that isolates RPE, itself. In other words, measured brain activity might either a) reveal the FRN to be an all-purpose RPE-computer, or b) ascribe the computation of RPE to various different neural sources, or c) reveal that RPE may not necessarily be computed in pure form, as tasks become complex, but instead, RPE may be modulated by particular other characteristics of the task.

Recently, Arbel, Murphy, and Donchin (2014) used a paradigm that is close to the current

task. On each trial, participants were presented with a novel object and had to choose from two non-words, one of which was the correct associate of the object. There were a total of 60 object–non-word pairs to be learned. Interestingly, the FRN for the correct feedback trials in cycle 1 was larger when followed by correct responses in cycle 4 than when followed by incorrect responses in cycle 4. On the other hand, FRN for the incorrect feedback trials in cycle 1 showed no such effect. The FRN-RPE account would have predicted that errors followed by correct responses would produce a larger FRN than errors followed by errors. Thus, it is possible that the FRN-indexed learning is more closely linked to how the feedback is used, rather than the RPE-based learning predictions, specifically when the task deviates from the commonly used gambling or time estimation tasks. Based on these findings, we also tested if and how the FRN indexed learning in the current task, by measuring this signal as a function of subsequent response adjustments. Specifically, for learning after surprise reversal, we hypothesized that the FRN for the switched trials, which offered more of a learning opportunity than the non-switched trials, will be larger when followed by correct responses than when followed by incorrect responses subsequently.

In sum, we re-evaluate the role of the FRN as a reinforcement learning error signal for a non-traditional RL task, where predicting reward is not the primary task, but rather, a guide to stimulus-specific learning.

4.2 Methods

4.2.1 Participants

A total of 68 (aged 19–22 years) introductory psychology students at the University of Alberta participated for the partial fulfillment of course credit. Data from 5 participants could not be retrieved due to machine error. Data from another 5 participants were excluded: 3 due to poor EEG signal quality, 2 due to many missing EEG triggers. Thus data from a total of $N = 58$ participants were included in the current experiment. All participants were required to have English as their first language and had normal to corrected-normal vision. Written consent was obtained prior to the experiment in accordance with a University of Alberta ethical review board. Prior to the experiment, participants were informed that the experiment was a "word choice task", and that they would receive a payment proportional to the total points earned in the experiment, in addition to their partial course credit.

Participants were informed that they could earn up to $5.00.

4.2.2 Materials

Stimuli were words selected from the MRC Psycholinguistic database (Wilson, 1988). Imageability and word frequency were all held at mid-levels and all words had six to seven letters and exactly two syllables. We additionally used the Affective Norms of English Words (Bradley & Lang, 1999) to exclude words with moderately arousing, positive, or negative emotional connotations which could interfere with learning the reward values (also used by Madan, Fujiwara, Gerson, & Caplan, 2012; Chakravarty et al., 2019). The final word pool consisted of 48 items. Half of the words were randomly chosen to be high-value and half to be low-value. Item values were randomized across participants.

4.2.3 Experimental Paradigm

The session took place in an electrically shielded, sound-attenuated chamber. Participants were told that they will be making word choices that will lead to rewards (see below and Figure 4.1). Participants were also told that they were to complete 19 cycles of choice trials following which they will be paid based on their total reward points accumulated. They were informed of their progress in the experiment along with the number of reward points accumulated after each cycle. The experiment consisted of two stages, unbeknownst to the participants: training cycles and surprise reversal.

Training cycles

On each trial (see Figure 4.1), participants were shown a word along with a non-word letter string, "HHHHH", on the computer screen simultaneously. The position of the word (left or right) was counter-balanced across all trials. All 48 words were presented once per cycle, and for a total of 19 cycles. Participants were instructed to either choose the word or the "HHHHH" by pressing the "P" or "Q" key of a computer keyboard to choose the item presented on the right or left side of the computer screen respectively. Responses led to either a high-value reward of 10 points or a low-value reward of 1 point. At the beginning of the experiment, half of the 48 words were randomly selected to be of high-value and the remaining to be of low-value. The participants were told that a high-value word would earn them 10 points if they chose it, or 1 point if they did not choose it (i.e., chose the "HHHHH"

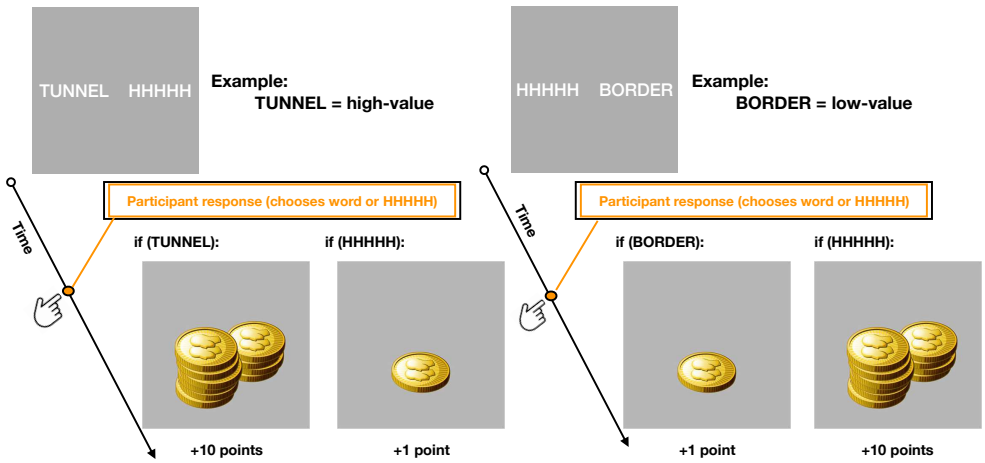

Figure 4.1: Illustration of a trial in the task. For high-value words, choosing the word led to the high (10 points) reward whereas choosing 'HHHHH' led to the low (1 point) reward. For low-value words, choosing the word led to the 1 point reward and choosing the 'HHHHH' led to the 10 points reward.

option). Conversely, a low-value word would earn them 1 point if they chose it, or 10 points if they did not choose it. Importantly, the participant could earn the 10 points reward for both high- and low-value words, depending on their choice. Thus, the two types of words were equally important in maximizing the total rewards earned. Aside from the verbal description of high and low-value, these two types of trials, theoretically, only differed in their response rules: high-value words required congruent responses, i.e., press the key on the same side as the word and low-value words required incongruent responses, i.e., press the key on the opposite side as the word. Moreover, using a similar reward learning paradigm, Chakravarty et al. (2019) showed that value did not influence memory when the high- and low-value words did not directly compete with each other. Trial choices were pseudo randomly generated, with each word used one time per choice set, but each choice set always consisted of one word and the string, "HHHHH". The word and letter string remained on the computer screen until the participant made a response. After each response, a 500–700 ms jittered interval followed before the presentation of the reward feedback at the centre of the screen for 1500 ms. The jitter introduced an uncertainty for when the feedback was expected to be seen by the participants. If they earned 10 points, an image of a pile of coins was presented; if they earned 1 point, an image of one coin was presented (Figure 4.1). The participant's current point balance was presented at the top of the screen during the feedback presentation. There was an inter-trial interval of 500–700 ms before the next choice trial started.

Surprise Reversal

The training cycles (16 cycles) were followed by a surprise reversal learning, which consisted of 3 cycles. In the reversal learning, the value of half of the words (both high- and low-value) were switched. This means, out of the 24 previously high-value words, 12 were randomly reversed to be of low-value (high-switched) and out of the 24 previously low-value words, 12 were randomly reversed to be high (or low-switched). Participants were not instructed about this change. This was a continuation of the training cycles and all items were presented in the exact similar manner. At the end of the reversal learning, which marked the end of the experiment, participants were paid using the conversion rate of $0.0006 for every point earned, rounded up to the nearest 25-cent amount. Participants earned between $3.00 and $5.00.

4.2.4 EEG recording

Scalp electrical activity was recorded using a high-density 256-channel Geodesic Sensor Net (Electrical Geodesics Inc., Eugene, OR), amplified at a gain of 1000 and sampled at 500 Hz. Impedance of each electrode was kept below 50 kΩ and the vertex electrode Cz was used as the reference. EEG signal was average re-referenced, and digitally bandpass filtered between 0.1–40 Hz. Data were analyzed by custom MATLAB scripts in conjunction with the open-source EEGLAB toolbox (`http://sccn.ucsd.edu/eeglab`; Delorme & Makeig, 2004). Artifacts such as eye blinks, muscle noise etc. were detected via Independent Component Analysis (ICA), implemented in EEGLAB. Trials were then epoched from 200 ms before to 1000 ms after the onset of the feedback stimuli. Baseline was removed. To detect epochs with possible artifacts, we used an absolute voltage threshold of 200 μV. For the same purpose, we also calculated the point-to-point difference between the time samples for each epoch and those with a point-to-point difference exceeding 25 μV were also removed. On average, 1.2% epochs were rejected.

4.2.5 Data Analysis

Based on previous studies (for a review, see Walsh & Anderson, 2012), FRN amplitude was computed from the fronto-central electrode FCz by averaging the signal over the window of 200–350 ms post feedback onset. Results were considered statistically significant based on alpha level of 0.05, but trend effects ($0.05 < p < 0.10$) are also discussed. We used SPSS (version 20) to conduct the statistical tests.

4.3 Results

We first report behavioural data, with special attention to the presence of two subgroups of participants identified by their pattern of behaviour during the reversal cycle 17. Then we report analyses of the FRN, taking these subgroups into account.

4.3.1 Behaviour

Accuracy Performance accuracy was measured by the proportion of trials in which the participant won the 10 points reward (Figure 4.2a). During the initial training cycles, performance accuracy differed between the high- and low-value words. However, initial perfor-

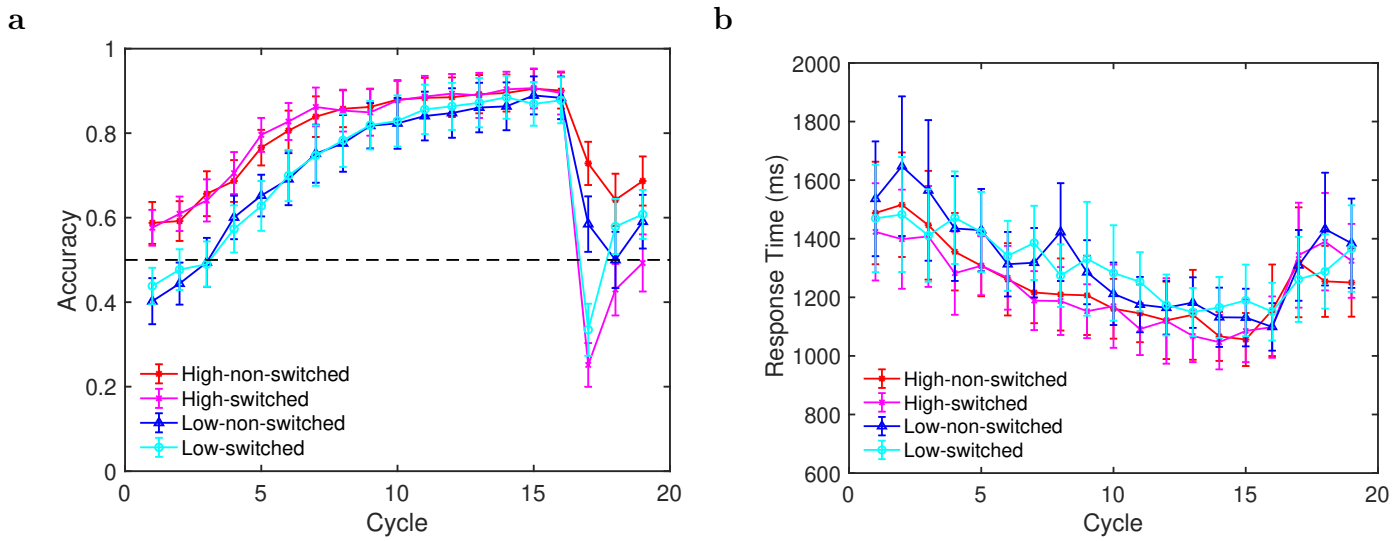

Figure 4.2: Learning curves for the training cycles. (a) plotting accuracy of responses (choose high or not-choose low); the dashed horizontal line indicates chance performance (0.5), and (b) response times for correct responses. All error bars represent 95% confidence intervals for the mean.

mance accuracy for words that later switched its values in the surprise reversal cycle was not different from that of the non-switched words. A 2×2 repeated measures ANOVA with the within-subject factors Value (high and low) and Reversal-state (switched or non-switched in cycle 17), on performance accuracy of cycle 1 showed a significant main effect of the Value, $F(1, 57) = 19.66,\ p < 0.001,\ MSE = 0.08,\ \eta_p^2 = 0.26$; accuracy for high-value words was significantly greater than that of low-value words, [mean±SEM, high: 0.58 ± 0.02, low: 0.42 ± 0.02]. Since participants had no prior knowledge about the values at this stage, this difference suggests that participants tended to choose the words more often than the non-word letter string "HHHHH", i.e., exhibited a word-choice bias. The other effects were not significant, which reassuringly indicated that there was no sampling bias between later-switched and later-non-switched words. The word-choice bias was greatly reduced as learning took place; an ANOVA with the same design, on the 16th cycle, showed no main effect of Value, $F(1, 57) = 1.64,\ p = 0.20,\ MSE = 0.01,\ \eta_p^2 = 0.03$ [high: 0.90 ± 0.02, low: 0.88 ± 0.03]. No other significant main or interaction effect was found either.

Response Times Figure 4.2b plots average response times for correct responses, as a function of cycle and the conditions mentioned above. Participants were in general faster in responding to a high-value word correctly than to a low-value word (also see Chakravarty

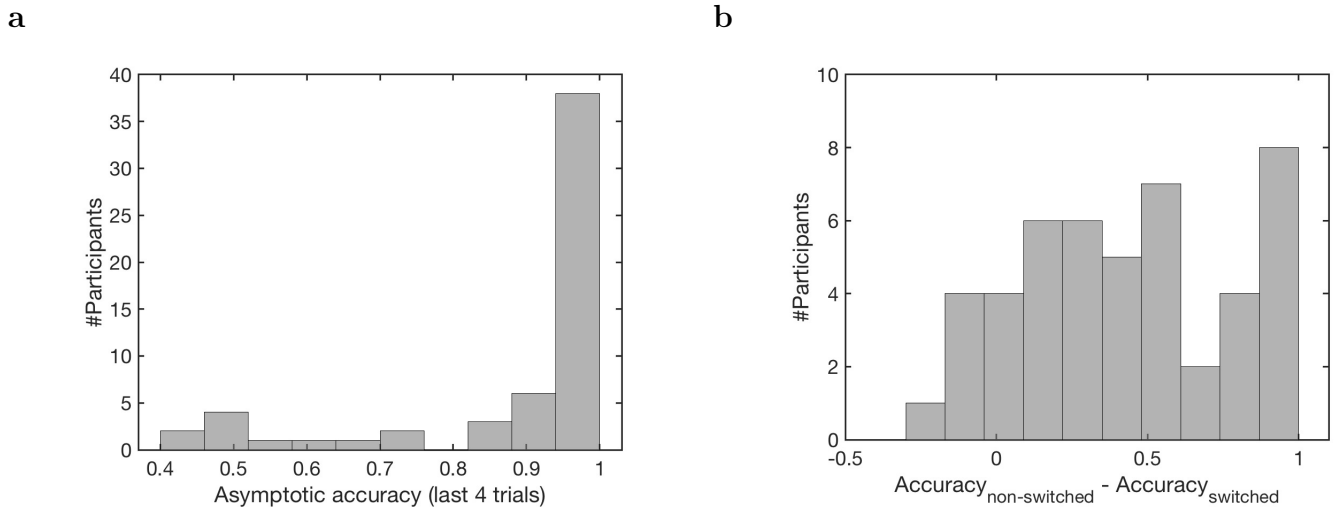

Figure 4.3: (a) Distribution of asymptotic accuracy— mean accuracy over the last four training cycles. (b) Distribution of the two strategies based on the accuracy difference between non-switched and switched trials in cycle 17.

et al., 2019), which could be due to the congruent and incongruent nature of responses required for them, respectively. A 2×2 repeated measures ANOVA with the within-subject factors Value (high and low) and Reversal-state (switched or non-switched) for the response times of cycle 1, revealed no main effect of Value, $F(1, 55) = 1.70,\ p = 0.20,\ MSE = 146804.48\ ms, \eta_p^2 = 0.03$, [high: $1441.86 \pm 84.63\ ms$, low: $1508.66 \pm 92.37\ ms$]. For cycle 16, there was an interaction between Value and Reversal-state, $F(1, 57) = 4.27,\ p < 0.05$, $MSE = 40688.29\ ms,\ \eta_p^2 = 0.07$; response times for low-non-switched ($1099.00 \pm 41.34\ ms$) tended to be faster ($p = 0.07$) than that of low-switched ($1151.09 \pm 50.32\ ms$), the trend appeared to be opposite ($p = 0.28$) for high-non-switched ($1155.71 \pm 79.76\ ms$) and high-switched ($1098.33 \pm 53.42\ ms$).

Overall, words chosen to switch values during the reversal learning were not found to be associated with pre-existing differences during the training cycles.

Non-learners The distribution of average accuracy over the last four cycles of value learning (i.e., cycles 13 to 16) showed that not all of our participants learned the values with equal competency (Figure 4.3a). This is important because if a participant is making random guesses about the values even after 16 cycles of learning, a surprise value reversal may not be detectable at all by that participant. We used a threshold of 0.8 to distinguish participants who showed good learning ($N = 47$) from those who did not ($N = 11$).

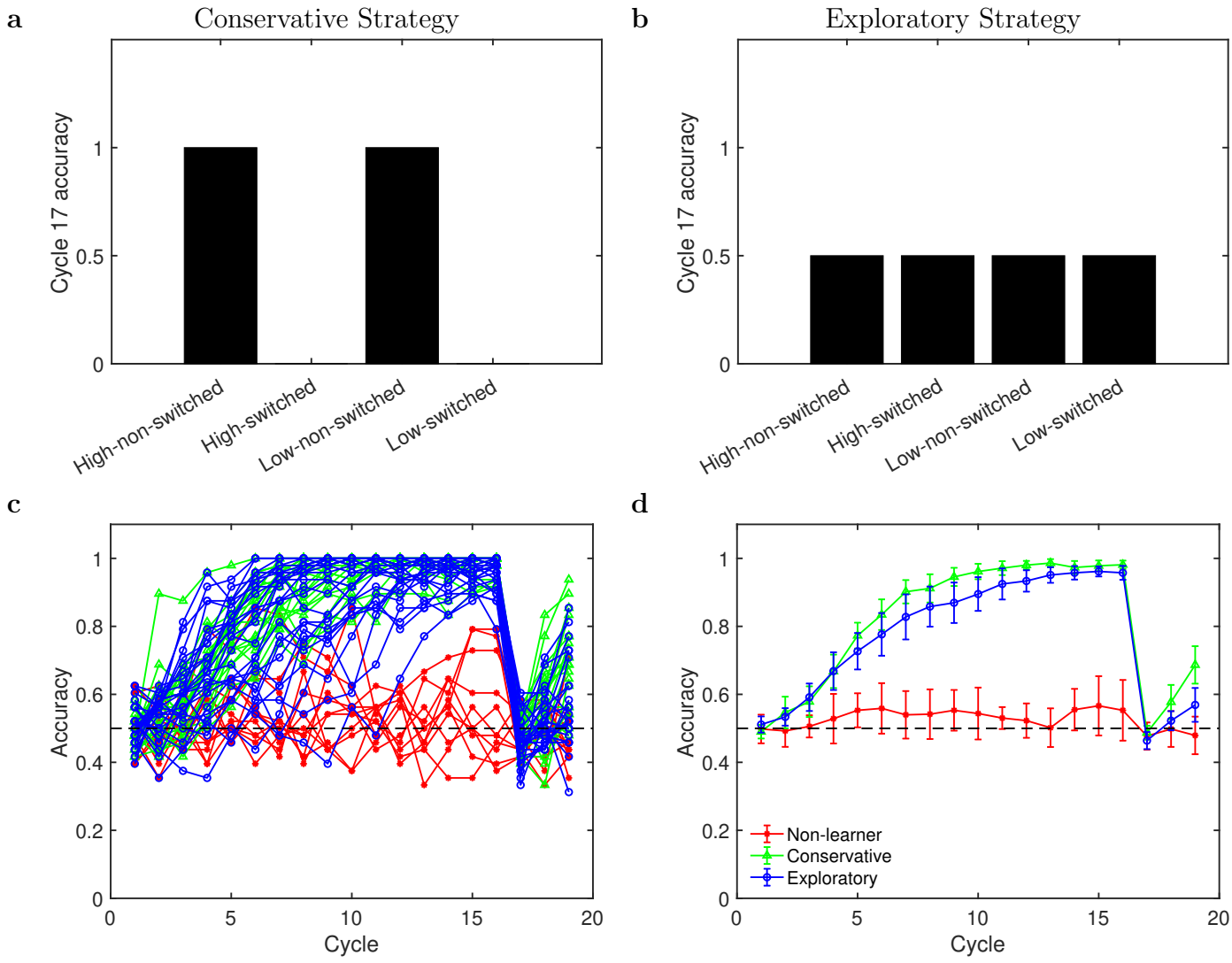

Figure 4.4: (a) Illustration of the extreme case for conservative strategy: correct on all non-switched trials, incorrect on all switched trials and (b) the exploratory strategy: accuracy at chance level for all non-switched and switched trials. (c) Learning accuracy for individual participants, separated by non-learners, conservative and exploratory strategies. (d) Mean accuracy for the training cycles. Note that both panels are averaged across the four conditions: high-non-switched, high-switched, low-non-switched and low-switched.

Two ways in which participants responded to the reversal cycle For participants who learned the values ($N = 47$), as expected, accuracy dropped close to chance level (50%) in cycle 17 when half of the word-values associations were reversed. Since the participants were not given any instruction about this value reversal in cycle 17, we had expected that they would respond correctly to the non-switched and incorrectly to the switched trials. However, we found that there were two ways participants responded to this change (Figure 4.4), which seem to be reasonable responses to unexpected feedback. At one extreme, the participant could choose to stick to their previous response choices about word-value associations (hereafter referred to as a "conservative strategy"), and at the other extreme the participant may choose to make random guesses, similar to the 1st cycle (hereafter referred to as an "exploratory strategy"). Participants who followed the conservative strategy (Figure 4.4a) would have made responses based on the word-value associations learned prior to the change (i.e., cycles 1 to 16). As a result, accuracy for non-switched words would be near perfect, whereas that for switched words would be close to zero. On the other hand, participants who followed the exploratory strategy (Figure 4.4b) would have made guess responses to the new word-value associations. As a result, accuracy for both non-switched and switched words would be close to chance level.

We derived an index of strategy by computing the accuracy difference between words that were not switched and those which were switched (Figure 4.3b). A strictly conservative participant would have an index of 1 (accuracy for non-switched = 1, switched = 0) and a strictly exploratory participant would have an index of 0 (accuracy for non-switched = 0.5, switched = 0.5). Thus, a threshold at the midpoint (0.5) was used to assign each participant to one of the two strategies. Participants with an accuracy difference above or equal to 0.5 were labelled conservative ($N = 21$) and participants with an accuracy difference below 0.5 were labelled exploratory ($N = 26$).

To check these observations, we conducted a 2×2×2 repeated-measures ANOVA with the between-subject factor Strategy (conservative and exploratory) and two within-subject factors Value (high or low) and Reversal-state (switched or non-switched), separately for the accuracy and response times and for cycle 1 and cycle 16. There were no significant main or interaction effect for Strategy for any of the 4 models. Thus, overall, there was little evidence for the quality of learning to be substantially different for the two strategies.

Reversal learning Following cycle 17, participants may have been able to improve, having partly adjusted to the changing rules. The learning curves show a rise in accuracy for cycles 18 and 19. We conducted a 2×2×2 ANOVA with within-subject factors Value and Reversal-state and between-subject factor Strategy on accuracy of cycle 18. This produced a significant interaction between Strategy and Reversal-state, $F(1, 45) = 5.74,\ p < 0.05,\ MSE = 0.15,\ \eta_p^2 = 0.11$. Follow-up tests showed that for both high-non-switched ($p < 0.05$) and low-non-switched ($p < 0.01$) trials, the conservative strategy (high-non-switched: 0.74±0.05; low-non-switched: 0.63±0.05) performed significantly better than the exploratory strategy (high-non-switched: 0.57±0.05; low-non-switched: 0.42±0.05). For switched trials, accuracy did not significantly differ between conservative (high-switched: 0.40 ± 0.06; low-switched: 0.54±0.06) and exploratory (high-switched: 0.45±0.05; low-switched: 0.66±0.05) strategies.

The interaction between Value and Reversal-state was significant too, $F(1, 45) = 22.32,\ p < 0.001,\ MSE = 0.05,\ \eta_p^2 = 0.33$. Follow-up tests showed that accuracy for high-non-switched (0.65 ± 0.04) was significantly greater ($p < 0.01$) than that for high-switched (0.43 ± 0.04); high-non-switched was also significantly greater ($p < 0.001$) than low-non-switched (0.52 ± 0.04); and accuracy for high-switched (0.43 ± 0.04) was significantly less ($p < 0.001$) than that for low-switched (0.60 ± 0.04).

The interaction Value×Reversal-state was also significant for cycle 19, $F(1, 45) = 9.88,\ p < 0.005,\ MSE = 0.05,\ \eta_p^2 = 0.18$. Once again, follow-up tests showed that high-non-switched (0.73 ± 0.03) accuracy was greater ($p < 0.005$) than high-switched (0.52 ± 0.04), high-non-switched accuracy was also greater ($p < 0.05$) than low-non-switched (0.63±0.04), low-switched (0.64±0.03) accuracy was greater ($p < 0.005$) than high-switched, and low-non-switched accuracy was greater ($p < 0.05$) than high-switched. Also, there was a main effect of Strategy, $F(1, 45) = 9.62$, $p < 0.005$ $MSE = 0.07$, $\eta_p^2 = 0.18$; conservative (0.68 ± 0.03) achieved higher accuracy than exploratory (0.57 ± 0.03). The main effect of Reversal-state was also significant, $F(1, 45) = 5.73,\ p < 0.05,\ MSE = 0.08,\ \eta_p^2 = 0.11$; non-switched accuracy (0.68 ± 0.03) was greater than that for switched (0.58 ± 0.03).

Thus, overall the two strategies elicited during the value reversal were not accompanied with significant differences during the initial training cycles. We saw a difference due to Value in handling non-switched versus switched words after the reversal, which is possibly due to the re-occurrence of the initial word-choice bias in response to learning the word-

value re-map. Further, participants with the conservative strategy performed significantly better than the exploratory group for cycle 19. However, the groups did not differ in prior learning rate; thus, conservative versus exploratory subgroups cannot be simply understood as stronger versus weaker participants. Due to their apparent qualitatively different approach to the task, and due to the fact that the two strategies sort trials differently across all possible conditions (see below), the first set of ERP analyses will be followed up with more fine-grained analyses that take these subgroups into account.

4.3.2 ERPs

Learners and non-learners Figure 4.5(a-b) presents the grand averaged ERPs for cycle 17 at the fronto-central electrode FCz, separated by learners ($N = 47$) and non-learners ($N = 11$) as well as by Reversal-state (non-switched and switched) and Value (high and low). For the time window of interest for the FRN (200–350 ms post feedback onset), there was a negative deflection for the learners. For the non-learners, this was not so clear. Figure 4.5(c-d) shows the average amplitude for this time window; for learners, for both high- and low-value words, switched trials were more negative than non-switched, which would be expected if switched trials produced larger RPE than non-switched. For non-learners, however, the trend appeared to be opposite– but the larger and overlapping error bars (SEM) suggest that this may not be conclusive. The scalp distribution of voltage for the difference-wave (switched − non-switched) for the FRN window showed a more frontal than mid-frontal negativity for the learners (Figure 4.5e); interestingly this was slightly oriented towards the right hemisphere in the case of the high-value words. We conducted a 2×2×2 repeated measures ANOVA on the mean FRN amplitude with the within-subject factors Value (high or low) and Reversal-state (switched or non-switched) and the between-subject factor Learner-status (learner or non-learner). There were no significant main or interaction effects. Learner-status×Reversal-state approached significance; $F(1, 56) = 3.86,\ p = 0.054$, $MSE = 4.39\ \mu V,\ \eta_p^2 = 0.06$. The form of this trend effect can be seen in Figure 4.5(c-d); for learners amplitude for switched ($-1.27 \pm 0.37\ \mu V$) was more negative than that for the non-switched ($-0.95 \pm 0.45\ \mu V$), whereas for the non-learners it went the opposite way (switched: $-1.09 \pm 0.77\ \mu V$; non-switched: $-2.14 \pm 0.93\ \mu V$). Thus, for the learners, a signal resembling the FRN in latency and polarity was present but it did not produce a significant difference between the switched and the non-switched trials. Also note that the scalp-topography of

this signal appeared to be more anterior than the typical fronto-central FRN topography. For the non-learners, it is difficult to interpret as to why there may be an opposite trend; one possibility is more noise in the data, specifically considering the large error bars. Another possibility is that since the non-learners were still making frequent incorrect responses at this point, by chance, this could have worked in favour of the switched than the non-switched, altering their predictions.

Conservative and Exploratory strategies Next, we looked into the ERPs for the two strategies, conservative and exploratory, separated by Value and Reversal-state (Figure 4.6). Non-learners were omitted from this and subsequent analyses. The negative deflection for the 200–350 ms time window post feedback onset was observed for both strategies (Figure 4.6a,b). Also, the mean amplitude over this time window appeared more negative for the switched than the non-switched, for both high- and low-value and for both conservative and exploratory strategies (Figure 4.6c,d). Amplitude of the FRN-like signal for the conservative participants was generally more negative than that for the exploratory participants (Figure 4.6c,d). However, this may not be specific to the FRN window, as Figure 4.6(a,b) shows that the ERP waveform for the conservative was overall further away from the baseline than that for the exploratory. The topographic plots for the difference wave (Figure 4.6e,f) showed frontal negativity for all conditions and once again, was slightly right frontal for the high-value trials. We carried out an ANOVA with the between-subject factor Strategy and the within-subject factors Value and Reversal-state. This produced no main effect of Reversal-state, $F(1, 45) = 2.14,\ p = 0.15,\ MSE = 2.37\ \mu V,\ \eta_p^2 = 0.04$. Reversal-state did not interact with the other factors. There was a main effect of Strategy; $F(1, 45) = 5.86,\ p < 0.05,\ MSE = 22.43\ \mu V,\ \eta_p^2 = 0.11$, exploratory: $-0.36 \pm 0.46\ \mu V$, conservative: $-2.04 \pm 0.52\ \mu V$. To check whether the difference due to Strategy was specific to the time window of the FRN-like signal, we computed the amplitudes for an earlier 50–150 ms window (FCz). Here too, the difference was close to significant; $F(1, 45) = 3.69,\ p = 0.06,\ MSE = 7.59\ \mu V,\ \eta_p^2 = 0.08$, exploratory: $-0.74 \pm 0.27\ \mu V$, conservative: $-1.52 \pm 0.30\ \mu V$. Thus, there may have been an overall difference in the waveform for the two groups, not just specific to the FRN-like signal. Overall, the comparisons between switched and non-switched, when broken down by the two strategies showed clearer trends than above but the FRN-RPE account was still not clearly present.

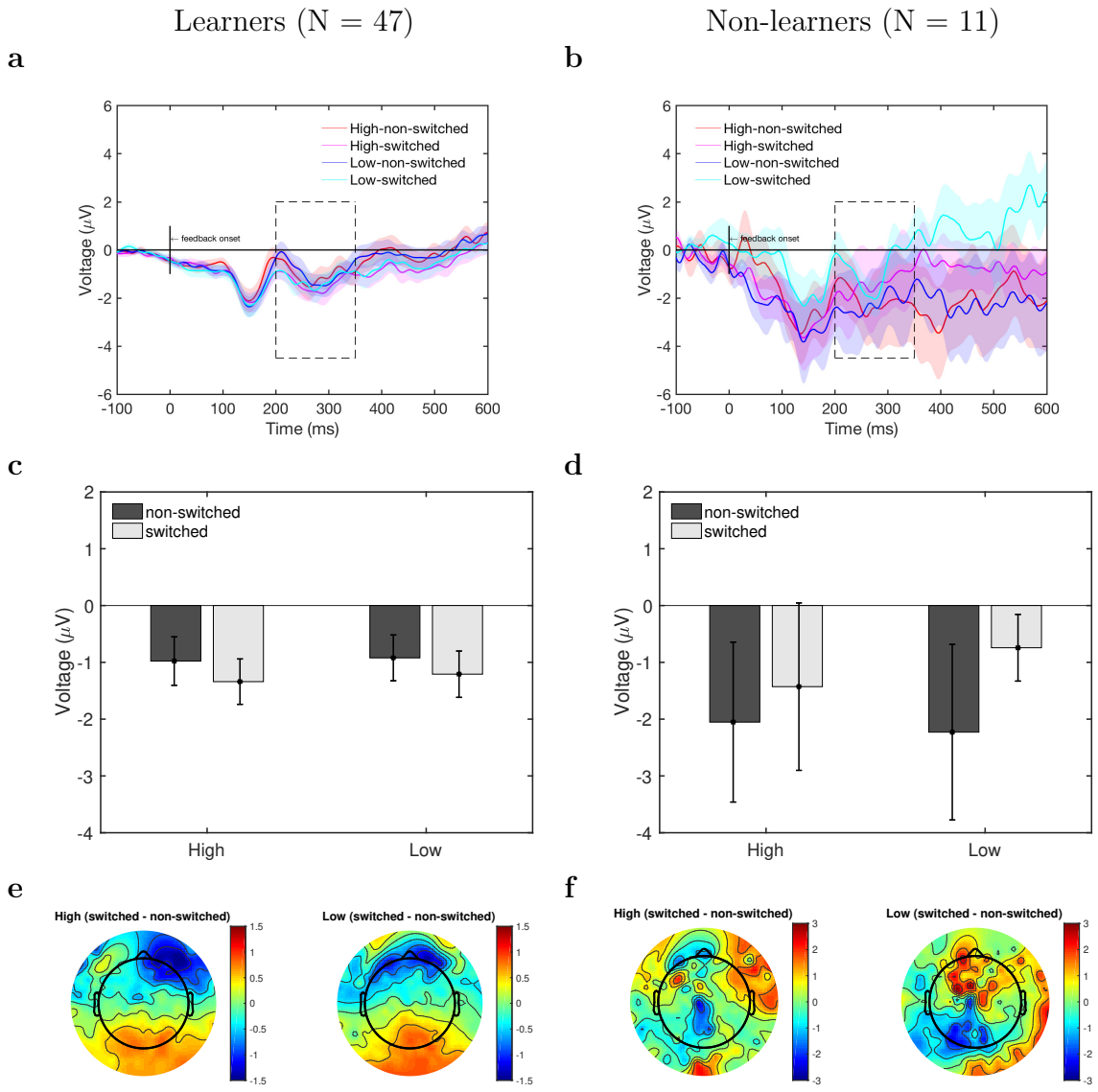

Figure 4.5: (a,b) Grand averaged ERPs at electrode FCz for cycle 17 and separately for learners and non-learners, broken down by value (high or low) and reversal state (non-switched or switched). (c,d) Mean amplitudes over the 200–350 ms time window post feedback onset. Error bars are standard error of the mean. (e,f) Scalp topographic plots of the difference wave (switched - non-switched) for the same time window (200–350 ms), color reflects mean voltage (μV). Note the color scale limits vary for learners and non-learners in the topographic plots.

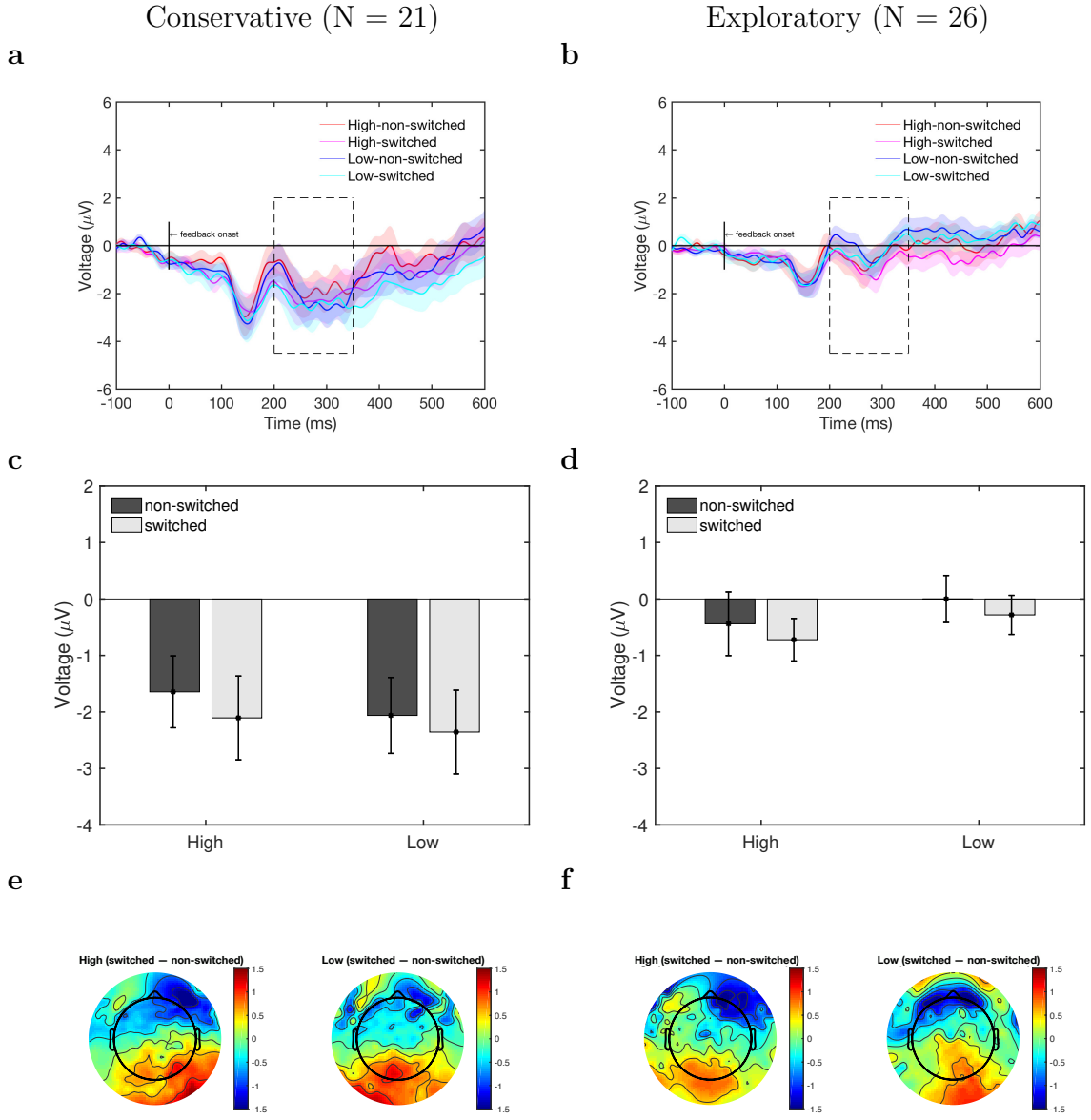

Figure 4.6: (a,b) Grand averaged ERPs at electrode FCz for cycle 17, separately for conservative and exploratory strategies and broken down by value (high or low) as well as reversal-state (non-switched or switched). (c,d) Mean amplitude over the 200–350 ms time window post feedback onset. Error bars are standard error of the mean. (e,f) Topographic plots of the difference wave (switched - non-switched) for the same time window, color reflects mean voltage (μV).

Response choices in cycle 17 (first reversal cycle) Our hypothesis for a larger FRN amplitude for the switched than the non-switched was based on the assumption that switched trials would lead to 'worse than expected' outcomes whereas for the non-switched, the outcomes will be 'same as expected'. However, if the participants frequently altered their responses in the wake of the reversal, as we found especially for the exploratory strategy participants, then the above assumption may not have always been true. In other words, if the participant altered their response, the corresponding prediction may have also changed. To test this, we re-evaluated the ERPs by further breaking them down into the two feedback outcome conditions (correct or 10 points reward and incorrect or 1 point reward). Then, we compared between switched and non-switched conditions for instances where the response was not altered. This means that for the non-switched condition, we chose trials where the outcome was correct (10 points) as this could have only happened if the response was the same. Likewise for the switched condition, we chose incorrect outcome trials, which could have only happened if the response was unaltered. We shall call this combined factor Feedback-outcome dependent Reversal-state. Thus, we compared non-switched-correct and switched-incorrect trials (Figure 4.7a-b). If the FRN-like signal observed in this task was tied more closely to the choice of responses in the reversal, then it may be possible to find a significant difference due to this signal for the non-switched-correct and switched-incorrect conditions. The mean amplitudes (Figure 4.7c-d) were overall more negative for switched-incorrect than non-switched-correct. The topographic plots of the difference wave (Figure 4.7e-f) showed similar trends as before. Once again, we carried out a 2×2×2 repeated measures ANOVA, with the within-subject factors Value and the combined factor Feedback-outcome dependent Reversal-state (non-switched-correct or switched-incorrect) and the between-subject factor Strategy. This showed a significant main effect for Feedback-outcome dependent Reversal-state; $F(1, 45) = 5.68,\ p < 0.05,\ MSE = 2.81\ \mu V,\ \eta_p^2 = 0.11$; switched-incorrect ($-1.57 \pm 0.36\ \mu V$) was more negative than non-switched-correct ($-0.99 \pm 0.35\ \mu V$). The main effect of Strategy also remained significant; $F(1, 45) = 4.21,\ p < 0.05,\ MSE = 20.69\ \mu V,\ \eta_p^2 = 0.09$, conservative: $-1.97 \pm 0.50\ \mu V$, exploratory: $-0.60 \pm 0.45\ \mu V$. All other effects were far from significant. Thus, taking response choices into account, we found a significant RPE-like effect.

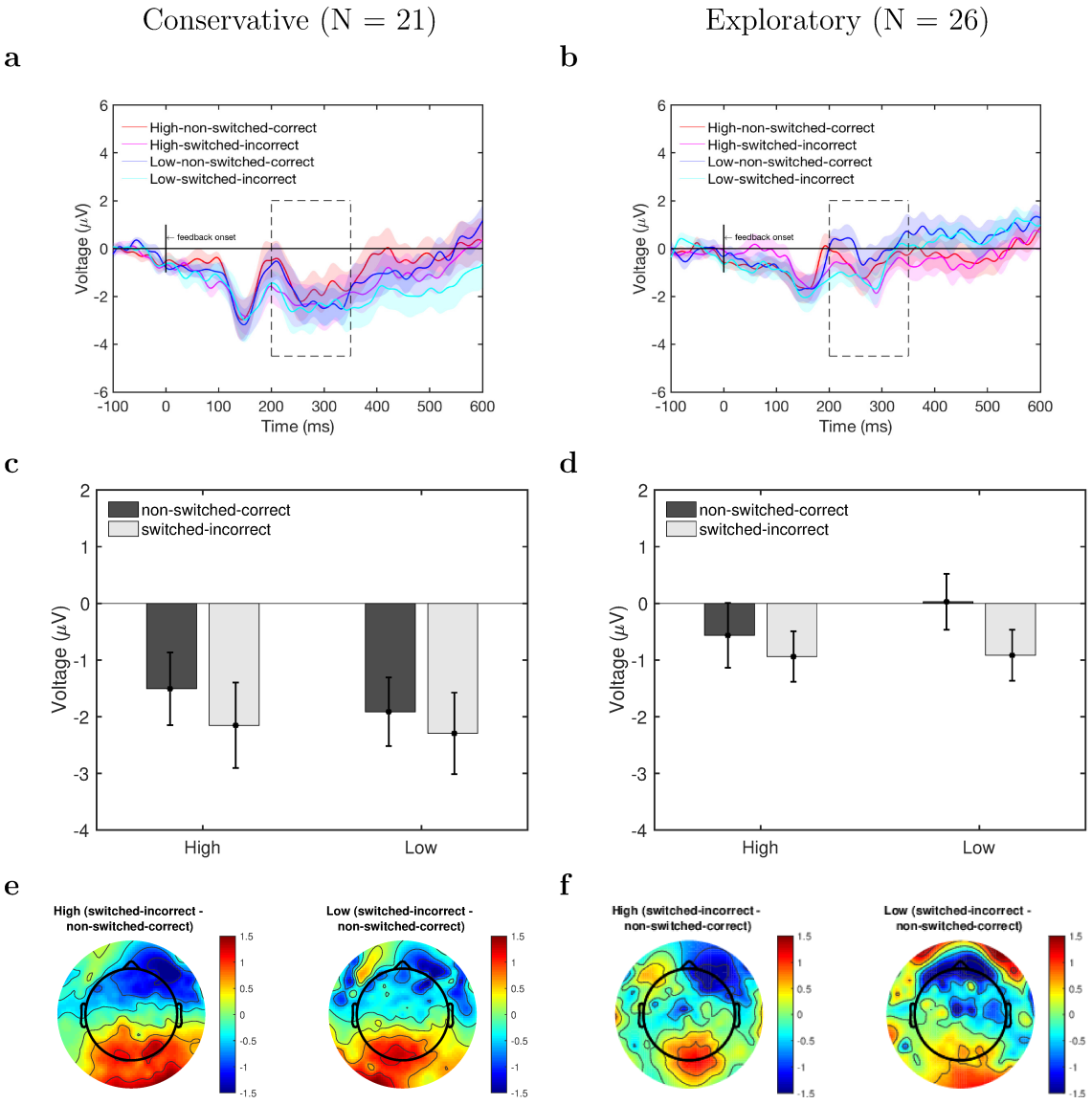

Figure 4.7: (a,b) Grand averaged ERPs at electrode FCz for cycle 17 and separately for conservative and exploratory strategies, broken down by value (high or low) and reversal-state combined with feedback-outcome (non-switched-correct or switched-incorrect). (c,d) Mean amplitudes over the 200–350 ms time window post feedback onset. Error bars are standard errors of the mean. (e,f) Topographic plots of the difference waves (switched - non-switched) for the same time window, color reflects mean voltage (μV).

Further analysis with the exploratory participants As mentioned before, participants following an exploratory strategy frequently altered their responses in the wake of the reversal. One possibility is that they reverted back to guessing the word-values. If this is true then on each trial, their prediction may have been halfway between the 1 and 10 points rewards, i.e., 5.5 points. Accordingly, correct responses (10 points reward) would have led to an RPE of +4.5 and incorrect responses (1 point reward) would have led to an RPE of −4.5. This would be true regardless of the value of the word (high/low) and reversal condition (switched or not). Figure 4.8 illustrates this hypothesis. Then, if the FRN indexed (signed) RPE for the exploratory strategy, we would expect a more negative FRN for incorrect than correct response trials, paralleling the hypothesized RPE in Figure 4.8. We tested this by breaking down the ERPs for cycle 17 into correct and incorrect feedback outcomes, for both switched and non-switched conditions.

This showed that for high-non-switched, incorrect feedback was actually more positive than correct (Figure 4.9a). For Low-non-switched, there was almost no difference between correct and incorrect (Figure 4.9d). For the switched conditions, however, incorrect feedback was more negative than correct (Figure 4.9c,f). The topographic plots (Figure 4.9b,e) for the difference wave (incorrect − correct) showed a frontal positive and negative signal for the high-non-switched and high-switched conditions respectively. Interestingly, the largest difference observed between correct and incorrect trials was that for low-switched and the corresponding topographic plot of the difference wave showed a clear fronto-central negativity. We conducted a 2×2×2 repeated measures ANOVA for the mean FRN amplitude with the within-subject factors Value (high or low), Feedback-outcome (correct or incorrect) and Reversal-state (switched or non-switched). This revealed a significant interaction for Value × Feedback-outcome, $F(1, 22) = 4.56$, $p < 0.05$, $MSE = 1.69\ \mu V$, $\eta_p^2 = 0.17$. Follow-up tests showed that low-switched-correct ($0.97 \pm 0.35\ \mu V$) was significantly more positive ($p < 0.05$) than high-non-switched-correct ($-.19 \pm 0.52\ \mu V$), high-switched-correct ($-0.42 \pm 0.64\ \mu V$) as well as high-switched-incorrect ($-0.89 \pm 0.50\ \mu V$). Low-switched-correct was also significantly more positive ($p < 0.005$) than low-switched incorrect ($-0.72 \pm 0.33\ \mu V$); also low-non-switched-correct ($0.36 \pm 0.45\ \mu V$) was more positive ($p < 0.05$) than low-switched-incorrect ($-0.72 \pm 0.32\ \mu V$). The main effect of Value approached significance; $F(1, 22) = 3.05$, $p = 0.09$, $MSE = 4.75\ \mu V$, $\eta_p^2 = 0.12$, high: $-0.36 \pm 0.40\ \mu V$, low: $-0.20 \pm 0.30\ \mu V$. The interaction Reversal-state×Feedback-outcome

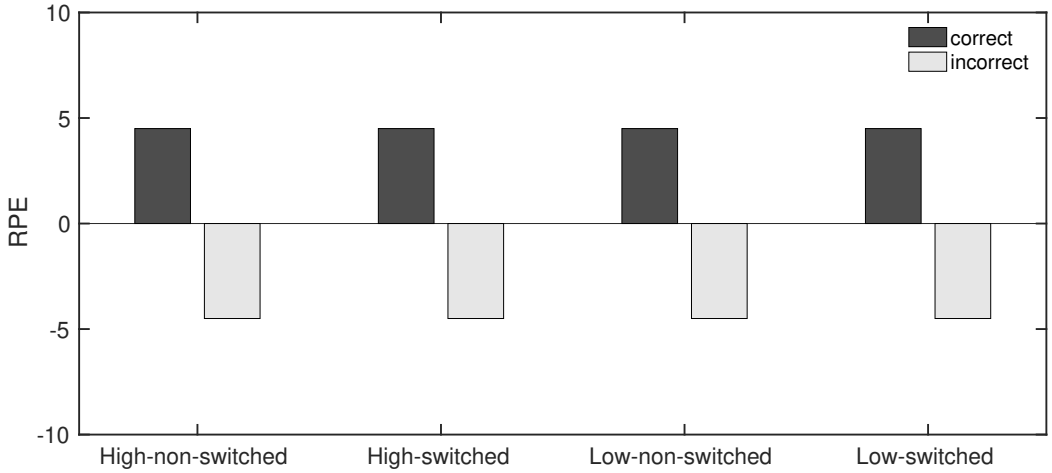

Figure 4.8: Hypothesized reward prediction errors for the exploratory strategy in cycle 17 if they had reverted back to guessing the word-values. We maintained the assumption that due to guessing they would have predicted a reward halfway between 1 and 10 points, i.e. 5.5 points. Then, based on whether they made a correct (10 points) or incorrect (1 point) response, the RPE should have been +4.5 or −4.5 respectively, across all value and reversal conditions.

also approached significance; $F(1, 22) = 3.42,\ p = 0.08,\ MSE = 4.38\ \mu V,\ \eta_p^2 = 0.13$, All other effects were far from significant. Taken together, these results show that the FRN-like signal was not overall more negative for incorrect feedback outcome than for correct feedback outcome, as would be expected if the FRN-like signal found in this task simply indexed an RPE-like function and if the exploratory strategy had simply reverted back to guessing in cycle 17.

FRN and response adjustments in subsequent trials If the FRN is a signal that guides learning, then one would expect that a large FRN amplitude would be predictive of later improvement in accuracy for a given item (as in van der Helden et al., 2010). We tested this possibility for the current task. Continuing with our analysis of the reversal trials, first we looked into ERPs from cycle 17, comparing trials that were subsequently responded correctly or incorrectly in cycle 18 (Figure 4.10).

For non-switched trials in cycle 17, across the two strategies and the two values (high/low), the FRN-like signal did not appear to drive the difference due to subsequent correct/incorrect responses in cycle 18 (see Figure 4.10a–b and the corresponding topographic plots for the difference wave). We conducted a 2×2×2 ANOVA with the within-subject factors Value

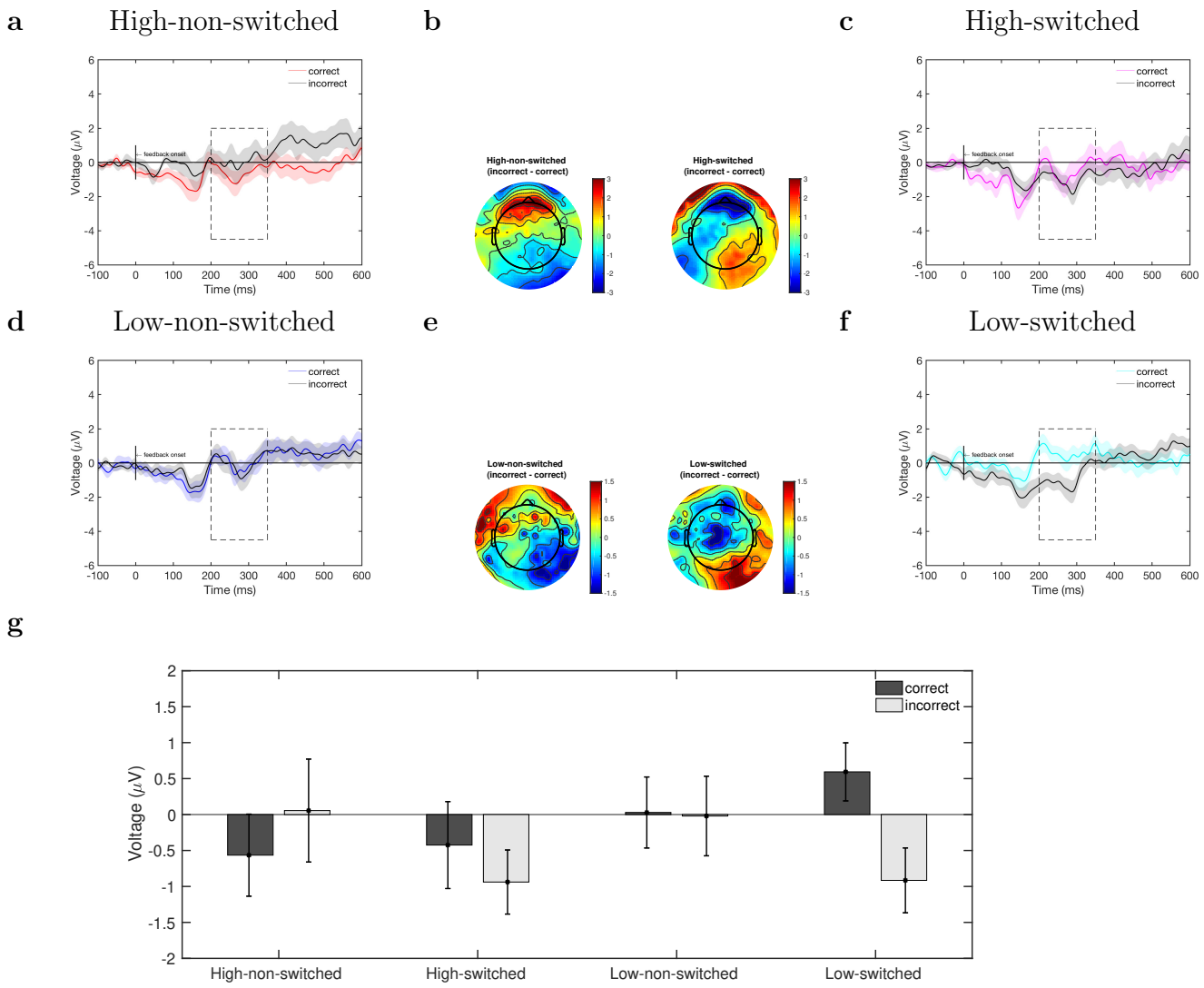

Figure 4.9: (a,c,d,f) Grand averaged ERPs at electrode FCz for cycle 17 and only for exploratory strategy, broken down by value (high or low), reversal-state (non-switched or switched) and feedback-outcome (correct or incorrect). (g) Mean amplitudes over the 200–350 ms time window post feedback onset. Error bars are standard errors of the mean. (b,e) Topographic plots of the difference waves (incorrect - correct) for the FRN time window, color reflects mean voltage (μV). Note the color scale varies for high- and low-value.

(high/low) and Subsequent accuracy (correct/incorrect) and the between-subject factor Strategy for the amplitude of the FRN-like signal for non-switched trials in cycle 17, which revealed no significant main or interaction effect for Subsequent accuracy or Value ($p > 0.05$). This could be because non-switched trials, in general, did not offer much of a learning opportunity.

For the switched trials in cycle 17, across conditions (see Figure 4.10c–d and the corresponding topographic plots) the FRN-like signal appeared to index difference due to Subsequent accuracy in cycle 18. We conducted another 2×2×2 ANOVA with the within-subject factors Value (high/low) and Subsequent accuracy (correct/incorrect) and the between-subject factor Strategy for the amplitude of the FRN-like signal for switched trials in cycle 17. This revealed a significant interaction for Subsequent accuracy×Strategy; $F(1, 38) = 4.84,\ p < 0.05,\ MSE = 5.21\ \mu V,\ \eta_p^2 = 0.11$. Follow-up paired t-tests showed that amplitude of the FRN-like signal for subsequently correct trials in cycle 18 ($-2.29 \pm 0.56\ \mu V$) was significantly more negative than that for subsequently incorrect trials ($-1.33 \pm 0.65\ \mu V$) for the conservative strategy (only for high-) $t(17) = 2.33,\ p < 0.05$. For the exploratory strategy, switched trials followed by correct responses ($-0.01 \pm 0.49\ \mu V$) were less negative than those followed by incorrect responses ($-0.65 \pm 0.56\ \mu V$) but the effect was not significant. Thus, adjustments to responses in cycle 18, based on the FRN-like signal (for the switched trials) in cycle 17, appeared to work differently for the two strategies.

Finally, we wondered if the FRN-like signal also indexed response adjustments in the early training cycles. To test this, we compared ERPs from cycle 1 for trials that were responded correctly or incorrectly in cycle 2. These ERPs were not broken down by Reversal state in cycle 17 as we did not expect to see any difference due to Reversal state in cycle 1 (also see the Behavioural results that confirm this result). Although the strategies were kept separate, we did not expect a difference in learning effects due to strategy because in cycle 1, participants using both strategies would have simply made guesses. For cycle 1, for the correct trials, and especially for the conservative strategy, subsequently correct responses in cycle 2 appeared to be more negative than incorrect responses in cycle 2 and the topographic plots of the difference wave (correct − incorrect) also supported an FRN-like signal (see Figure 4.11a–b). A 2×2×2 ANOVA with the within-subject factors Value (high/low), Subsequent accuracy (correct/incorrect) in cycle 2, and the between-subject factor Strategy, revealed a significant main effect for Subsequent accuracy; $F(1, 41) = 4.94,\ p < 0.05,\ MSE = 2.21\ \mu V,\ \eta_p^2 = 0.11,$

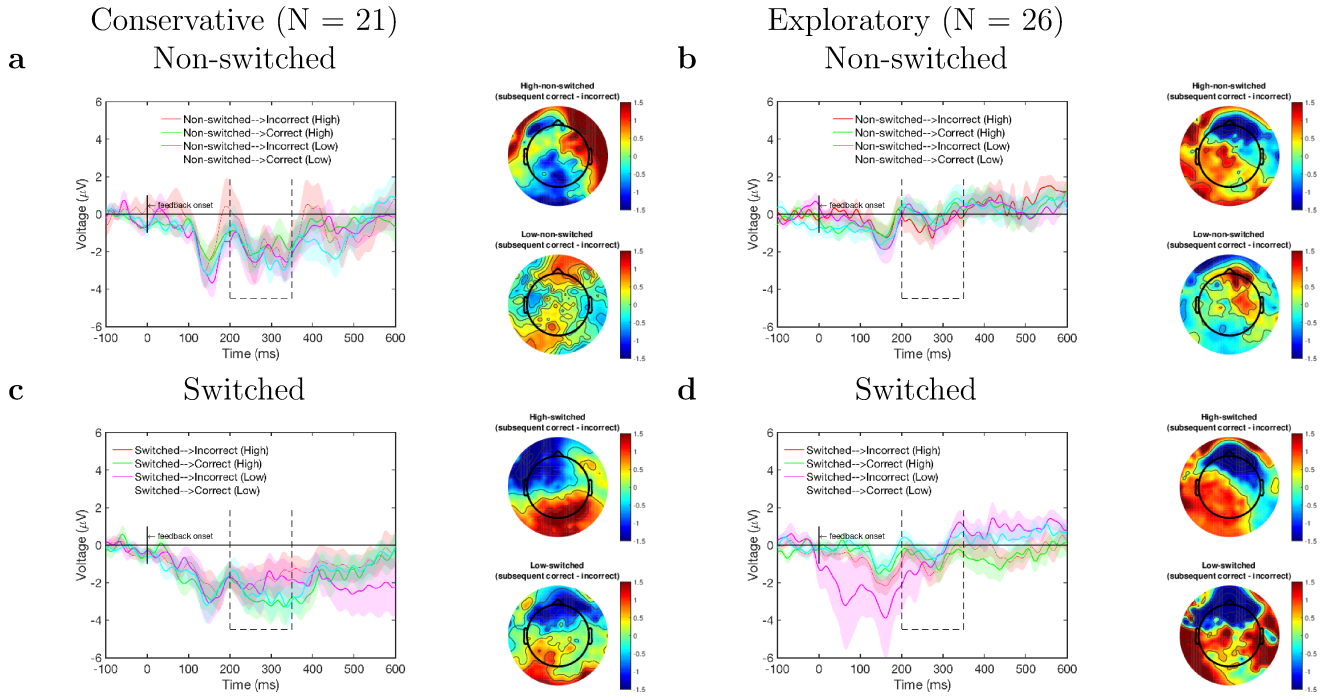

Figure 4.10: ERPs at FCz for cycle 17 for non-switched (a–b) and switched (c–d) trials that were followed by a correct or an incorrect response in cycle 18. Topographic plots of the difference wave (subsequent correct − incorrect) are placed next to the ERPs, color reflects mean voltage (μV) over the 200–350 ms time window post feedback onset. ERPs are broken down by conservative (left) and exploratory strategies (right) as well as by Value (high/low). Error bars are standard errors of the mean.

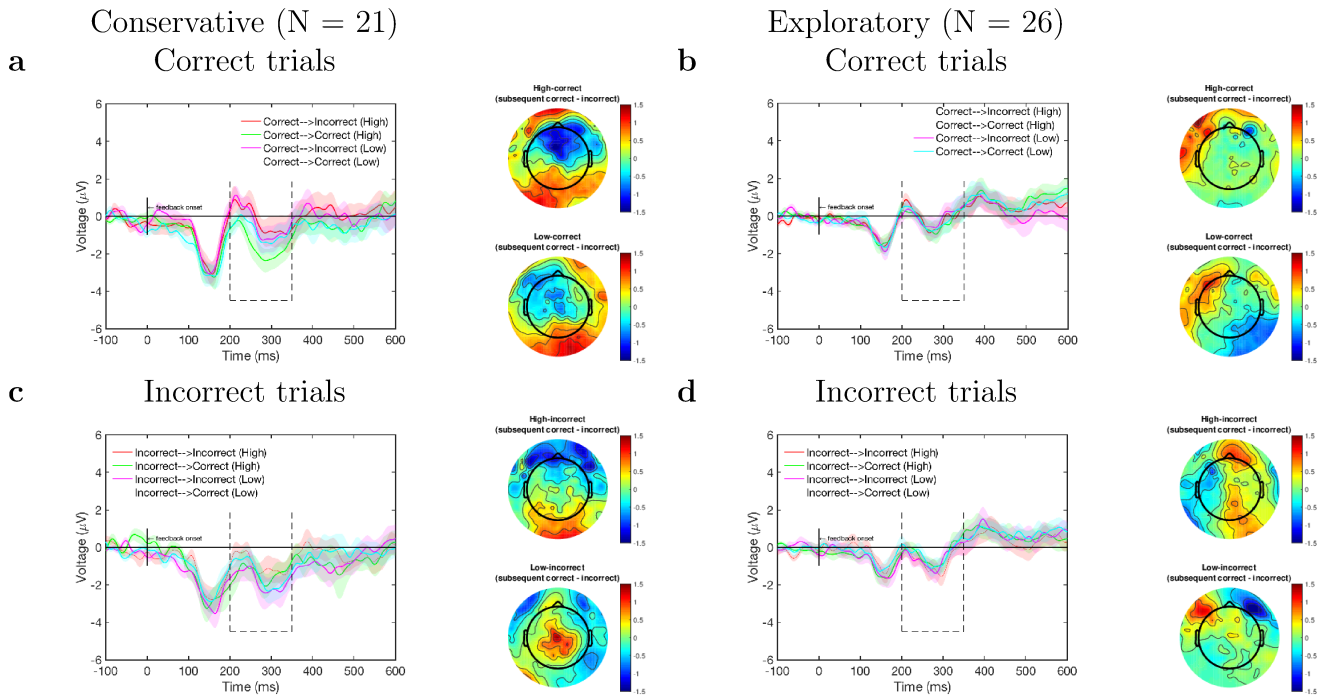

Figure 4.11: (a–d) ERPs at FCz for cycle 1, for correct and incorrect trials that were followed by a correct or an incorrect response in cycle 2. Topographic plots of the difference wave (subsequent correct − incorrect) are placed next to the ERPs, color reflects mean voltage (μV) over the 200–350 ms time window post feedback onset. All ERPs are broken down by conservative (left) and exploratory strategies (right) as well as by Value (high/low). All error bars are standard errors of the mean.

subsequently correct ($-0.63 \pm 0.29\ \mu V$) trials in cycle 2 were associated with an overall more negative amplitude for the FRN-like signal in cycle 1 than subsequently incorrect ($-0.12 \pm 0.34\ \mu V$) trials in cycle 2. For the incorrect trials in cycle 1 (see Figure 4.11c–d), a similar design ANOVA revealed no significant main or interaction effect.

Thus, the FRN-like signal did appear to index subsequent response adjustments in cycle 2 but only for the correct feedback outcome trials in cycle 1.

4.4 Discussion

We sought to test whether the role of the FRN in computing RPE generalized to a task for which RPE guided stimulus-specific learning. Here we discuss the full set of results with respect to this basic hypothesis. We had predicted a more negative FRN for feedback following the switched than the non-switched trials. A signal resembling the FRN did appear— a negative deflection in the window of 200–350 ms post feedback onset but with a more frontal

than mid-frontal negativity in the distribution of voltage across the scalp. Thus, it is possible that the signal evaluated here is different from the FRN signal suggested by previous studies. Moreover, this signal did not significantly differentiate between switched and non-switched conditions; despite participants having learned word-values to a high accuracy criterion prior to the reversal trials.

Individual variability in behaviour during the surprise reversal further informed this basic result. If participants continued to respond based on their previous knowledge about the word-value mappings (i.e., prior to the reversal), they would be incorrect for most of the switched and correct for most of the non-switched trials. Therefore, the absence of modulation of the amplitude of the FRN-like signal between these two conditions would indicate that the FRN-like signal did not index RPE-like effect in our task. Although this logic applies to one subset of participants (those using the "conservative" strategy), a second set of participants appeared to make new guesses during the reversal block (those using the "exploratory" strategy). Thus, it is possible that whenever the participant chose to alter their response (i.e., make a guess) their prediction for that trial was also different from when they did not alter the response. To factor this into the analyses, we compared between non-switched and switched conditions for instances where the response was not altered, i.e., non-switched trials that were met with a correct feedback outcome and switched trials that were met with an incorrect feedback outcome. In this case, the FRN-like signal was significant, with more negative amplitude for switched-incorrect than for non-switched-correct. This can be viewed as a conceptual replication of the FRN-RPE mapping.

One might infer that the exploratory and conservative participants simply differed in their learning ability, but behavioural results showed no significant difference in accuracy between the two groups, either at the beginning of the training cycles or just before the reversal. Rather, a significant difference between the two strategies appeared when looking into the question of how the FRN-like signal may have guided learning after the surprise reversal. Consider that in computational RL, the RPE for a trial is added back to the next prediction (multiplied by the learning rate); the updated reward prediction is used to determine the next response. Accordingly, if the FRN indexes RPE, larger deflections of the FRN could index response adjustments in the next trial. To test this, we looked into the amplitude of the FRN-like signal for cycle 17 trials that were followed by correct responses in cycle 18 and those followed by incorrect responses in cycle 18. When restricted to non-switched trials in

cycle 17, there were no effect on its amplitude due to subsequent response accuracy in cycle 17. This may be expected since for the non-switched trials, responses were overall more correct and there was not much to learn. For the switched trials in cycle 17, the FRN-like signal was significantly more negative for subsequently correct than incorrect responses, as we would have predicted, but only for the conservative strategy. For the exploratory strategy, the effect was opposite and non-significant.

One possibility is that since the conservative strategy did not alter their responses as frequently as the exploratory, they might have followed the feedback more closely to learn the changes, which produced the significant learning effect for this group as indexed by the FRN-like signal. The exploratory strategy participants, on the other hand, had to keep track of their responses as well as process the feedback to learn the changes, though we can not determine the relative amount of dependence on the response and the feedback signal for this strategy. Regardless of the correct view, these lines of thought suggest that learning indexed by the FRN-like signal account could depend on the way the feedback signal is used (for an individual-differences approach to this question, see Arbel & Wu, 2016).

Reassuringly, the difference due to the two strategies in learning as guided by the FRN-like signal was not present when considering trials from the very first training cycle, which would be expected as at this stage both groups were simply guessing. Interestingly, the FRN-like signal in cycle 1, for correct feedback trials, was more negative when followed by correct responses in cycle 2 than when followed by incorrect responses in cycle 2. For the FRN-like signal in cycle 1, when restricted to the incorrect feedback trials, no such effect due to subsequent response accuracy in cycle 2 was found. This is in line with the finding of Arbel et al. (2014), who, for a paired associate learning task, found similar effects for the positive and not the negative feedback trials in cycle 1 in connection with subsequent response accuracy in cycle 4.

Consider that in our task (as well as in Arbel et al., 2014), on each trial there were only two options to choose from, one of which was the correct response. Thus, correct and incorrect trials were equally useful in determining the response for the subsequent trial. Based on the FRN-RPE account, we would have predicted that errors followed by correct responses would produce a larger FRN than errors followed by errors. However, on cycle 1, both correct and incorrect outcomes are due to random guesses. Further, due to large number of stimuli, we could expect interference in memory for the stimulus-response mappings. Thus,

the significant effect for the FRN-like signal for the correct trials only (in cycle 1, as functions of response accuracy in cycle 2) could also suggest that the FRN-driven learning effect is more specific than RPE-driven learning proposed by RL theories.

During the surprise reversal, the responses, at least for the conservative strategy participants, were not all random guesses. Thus, this group of participants would have clearly more to learn from the switched than the non-switched trials. Accordingly, the FRN-like signal indexed learning for the switched and not the non-switched and for the conservative strategy only. In contrast, Ernst and Steinhauser (2012) found that while learning to associate Swahili words with German words, FRN amplitude (with the peak-to-peak measure) for error trials in cycle 1 that were followed by error responses subsequently was more negative than those followed by correct responses, which is similar to the trend (although non-significant) we saw for the exploratory strategy for the switched trials in cycle 17 as a function of response accuracy in cycle 18. Taken together, these results cast doubt on the role of the FRN being limited to a generic reward prediction error only. Instead, the FRN could be indexing learning effects based on task specific cognitive demands.

As mentioned above, our results challenge the idea that the FRN may play a broader role in learning. Instead, the FRN-like signal may be situated more locally within the task. Some have also found that there may be specific, rather than generic, connections between the FRN and a spillover effect on subsequent memory success. For example, in a study by Arbel, Goforth, and Donchin (2013), participants learned to associate novel objects with non-words (four response options to choose from, on each trial) and with feedback, for a total of 30 associations, each repeated 20 times. On the following day, the participants came back and were provided with all the objects and words and asked to match them. For the correct feedback trials, associates that were later recognized in the test had a more negative FRN than those not recognized. No such relation was observable for the negative feedback trials. Höltje and Mecklinger (2018) also showed that in a task where participants learned to pair Chinese characters with arbitrary images (which were also the feedback) and was followed up with a memory test for the feedback images, the FRN for the positive feedback images was larger when the feedback image was subsequently remembered versus forgotten.

Collins and Frank (2018) took an interesting approach to examine possible connections between RL and working memory. Rewards were deterministic and the participant learned to choose the correct response for each stimulus through repetition. Across different

blocks/cycles the Stimulus set size were varied (maximum = 6), as well as delay (number of trials since last correct), as ways of manipulating working memory demand. A Q-learning model computed expected reward (Q value) and RPE for each trial. For the stimulus-locked EEG, effects (β weights) of Q were strong for the 200–400 ms time window and significant for electrodes in central as well as posterior regions, whereas for number of stimuli and delay, effects were within the 300–700 ms and 200–600 ms windows respectively. Thus, both RL and factors relevant to working-memory predicted voltage. For the feedback-locked signal, RPE effects were found near the 200–400 ms window but the corresponding significant electrodes were located in the central-posterior scalp (e.g., Cz, CPz), which is consistent with the latency but not the topographic map of the FRN. However, for a later window near 400–600 ms, the RPE effects were located in the fronto-central area (e.g., FCz, Cz), which is consistent with the topographic map but not the latency of the FRN. In the feedback-locked signal, the effect of number of stimuli was weak but that for RPE×delay may have been present. Moreover, EEG correlates of both Q (for stimulus-locked) and RPE (for feedback-locked) increased with number of stimuli, suggesting that the RL and working-memory systems may not be independent. When the number of stimuli was small, EEG indices (voltage) of RPE declined much faster.

Our task was not designed to manipulate working memory and the number of stimuli used in this study was considerably higher than the maximum set size used by Collins and Frank (2018), though the number of exposures to each stimulus (maximum 15) was close to the number of exposures for our task (16). However, in light of their results, it is possible that the RPE in this task also declined at a much slower rate over the course of the training cycles. This could have contributed to a meaningful difference between the current paradigm and the two-armed-bandit and other similar FRN paradigms. Namely, the slower RPE function could also have been the case during reversal learning, translating into weaker FRN-like differences overall. There may have also been differences in the RPE decline rate due to the two strategies, paralleling the small difference in the slope of the learning curves for these two groups.

Turning to the word-value effect, recall that high- and low-value words only differed in the way these were described at the beginning of the experiment and how the response (in favour of the 10 points reward) had to be made. Aside from these, the utility for these two types of words in achieving the goal of maximizing the total rewards, was the same. Accordingly, we

do not expect the quality of memory for high- and low-value words to be different. Using the same basic training cycles but without the reversal cycles, Chakravarty et al. (2019) found no difference in subsequent free recall (with or without incentives) accuracy for the words. However, some difference due to value in lexical decision times was present, likely due to habituation of the congruent and incongruent actions involved in making correct responses for the high- and low-value words respectively, during the learning phase. The difference in response times was also observed in the current task. Moreover, difference due to value in learning accuracy for the cycle 1 can only be explained by a bias to choose the word than the 'HHHHH' string, because the participants could only make guesses at this stage. Additionally, it is possible that this bias reappeared, at least for the exploratory participants when they made guesses in the wake of the surprise reversal. The ERPs for cycle 17, for the exploratory strategy, showed a significant interaction between value and feedback accuracy, which was mainly driven by the low-switched and the high-non-switched conditions (feedback correct versus incorrect). The low-switched trials showed the largest FRN-like signal across all examined conditions. One possibility is that since low-switched trials were those instances where now choosing the (previously low) word led to better outcome than not-choosing it, the word choice bias would have led to the high reward for the low-switched trials. The word choice bias will also lead to the high reward for the high-non-switched trials, however here correct responses were when the participant did not alter their response and thus, may have been more surprised to receive the same outcome as before (i.e., with respect to the overall changed status). Thus, overall, it is possible that a word-choice bias in cycle 17 influenced how the trials were sorted for whether or not there was a violation of expectation.

Although post-hoc, features of the results suggest why the FRN-like signal may have retained the hypothesized role in indexing subsequent response adjustments only for the conservative participants (for the switched trials). The learning curves show that although both strategies start with similar performance and also achieve similar levels of performance by the end of the training cycles, there may be a small difference in the rate of learning as indicated by the slopes of the learning curves; with the exploratory strategy being slower. This difference was also observed in the last two cycles after the surprise reversal, with the conservative participants re-learning the word-values a little better than the exploratory. Thus, the conservative strategy may have been slightly more efficient for learning the reversals (only keep track of feedback and update response) than the exploratory (keep track

of both one's own responses and the corresponding feedback). Specifically, the conservative participants may have relied on RPE more exclusively than the exploratory participants. The RPE does not directly depend on the response of the current trial but does influence the response for the next by updating the reward expectation. Considering the putative FRN-RPE connection, our results accordingly showed no substantial difference in the FRN-like signal due to strategy for cycle 17 (current trial) but due to the responses in cycle 18 (future response). Thus, we show evidence for the FRN-like signal tracking learning, possibly when learning is more closely dependent on the RPE. On the other hand, our results also show that when a large number of stimuli are involved, the RPE-dependent RL mechanism could vary, potentially due to (differential) involvement of other cognitive functions, such as interference in memory for the stimulus-response pairs.

Relevant to the difference due to how the feedback signal is used for learning, Walsh and Anderson (2011a) instructed their participants about the reward probabilities of each stimulus, eliminating overt dependence on feedback to do well in the task. However, the FRN-RPE relevant effects were still present, though a difference in how the FRN indexed learning may have appeared. Yeung, Holroyd, and Cohen (2004) found differences due to FRN amplitude even when participants made no overt response and simply processed the reward feedback passively. Taken together, these findings may suggest that the functionality of the FRN is more extensive than that put forward by the FRN-RPE account. However, under specific learning situations, such as those present in a typical gambling/time estimation task, the FRN functions primarily like a reinforcement-learning error signal.

Our findings also appear consistent with Cockburn and Holroyd (2018), who found that as the information content of the feedback increased, such as by providing the participants with reward probability, magnitude, margin of error etc., the FRN was attenuated. Our exploratory strategy group may have only processed the feedback magnitude (correct or not) while the conservative strategy group used it to relate back to the preceding words. Thus, the same feedback may have afforded more information to the conservative than the exploratory participants, leading to the absence of the hypothesized effect of learning indexed by the FRN-like signal for the exploratory group.

In sum, our findings suggest that the FRN may not index RPE in all RL situations. It is possible that as the RL problem becomes more complex, RPE computations are influenced by other features of the task. Moreover, the topographic map of the FRN-like signal observed

in this task may suggest a different albeit close source compared to the FRN signal reported by previous studies. The FRN-like signal may not be computing an untarnished RPE value, but rather, computing more local RPE, specific to each stimulus, and relative to the response and possibly even the type of feedback just received.

Chapter 5

Predicting trial-and-error learning with brain activity

Abstract

In trial-and-error learning, feedback is used to update current knowledge. Thus, analysis of brain-activity during feedback-processing could identify signals that support trial-to-trial learning. Going beyond traditional analysis, which is based on planned-comparisons and descriptive methods, we asked if brain-activity during feedback-processing *predicted* trial-to-trial learning. Participants learned, through repetitions (cycles), the response-rules for a set of 48 words, divided equally into high- and low-value. The goal was to maximize the total rewards, which could be achieved by choosing the high- and not choosing the low-value words. We found that amplitude of an event-related potential signal resembling the latency of the feedback-related negativity (FRN), which is thought to index the discrepancy between expected and actual outcomes, or reward-prediction error (RPE; Holroyd & Coles, 2002), modestly predicted trial-to-trial learning, but only when the previous trial was correctly responded. Thus, different from RPE, which supports response adjustments following errors, this FRN-like signal may support maintaining of correct (learned) responses from one trial to the next. We also investigated multivariate pattern analysis of EEG-activity (time-domain) during feedback-processing, but due to the small number of trials available for training the classifiers, it was under-powered and thus, failed to predict trial-to-trial learning. Interestingly, classifier-performance increased with participants' performance. Additionally, multivariate pattern analysis of brain-activity during feedback-processing achieved modest success in predicting the inferred word-values (high/low); suggesting that participants may be 'thinking back' about the preceding stimulus after receiving the feedback. Also, the classifiers identified distinct patterns of brain activity for predicting word-value, when the feedback-outcome was correct than when it was incorrect, which suggested that learning from correct and incorrect feedback-outcomes may be supported by signals that are recruited differently. Together, these findings show promise in isolating behaviourally-relevant brain activity for trial-and-error learning. They also call for a reconsideration of the ways in

which reinforcement learning theory can inform our understanding of human trial-and-error learning.

5.1 Introduction

Many forms of learning are driven by feedback. A feedback signal informs us about the quality of the current response, so that future responses/actions can be adjusted in favour of the better outcomes. Thus, analysis of brain activity during feedback processing can identify signals that reflect cognitive processes driving learning from the feedback, our current aim. For electroencephalographic (EEG) recordings, an event-related potential (ERP) peak, namely, the feedback-related negativity (FRN), is likely elicited when processing a feedback-stimulus. It is commonly characterized by a mid-frontal negativity in the scalp-distribution of voltage, and has a peak-latency close to 300 ms post onset of the feedback-stimulus (Miltner et al., 1997; Nieuwenhuis et al., 2004). Across many studies, the FRN has shown a greater deflection for negative- than positive feedback outcomes (Bellebaum & Daum, 2008; Holroyd et al., 2004; Pfabigan et al., 2015; Marco-Pallares et al., 2011; Hajcak et al., 2006). Others have also found that amplitude of the FRN is modulated by differences in reward-magnitude (Yeung, Botvinick, & Cohen, 2004), reward-probability (Goyer et al., 2008; Nieuwenhuis et al., 2002; Walsh & Anderson, 2011b); for wins and losses (Frank et al., 2005; Yeung, Botvinick, & Cohen, 2004); as well as for rewards and punishments (Yeung, Botvinick, & Cohen, 2004; Wu & Zhou, 2009). Together, these findings have led to the suggestion that the FRN indexes discrepancies between expected and actual outcomes, also known as reward-prediction error (RPE; see Holroyd & Coles, 2002). Thus, similar to the RPE function, the FRN may also support trial-and-error learning (e.g., Cohen et al., 2007). However, the traditional ERP analysis of the FRN is based on planned-comparisons and descriptive methods, which, as we explain later in this section, can overestimate brain-activity measures by overfitting it, and also do not allow discovery of subtle multivariate pattern of activity that may be more relevant for explaining trial-and-error learning behaviour. In contrast, here we used a stronger, predictive framework to evaluate the role of the FRN as well as multivariate brain activity during feedback processing, to explain trial-and-error learning behaviour.

The RPE account of the FRN (Holroyd & Coles, 2002) is motivated from reinforcement learning theories (Sutton & Barto, 1998). In reinforcement learning, RPE is used to update reward-expectations for subsequent trials, depending on the learning-rate. There is evidence for RPE-like computations in the brain. For example, in animals, phasic dopamine activity

in the midbrain is found to code for RPE (Schultz et al., 1997). For better-than-expected outcomes (positive RPE), there is phasic increase in activity (relative to the spontaneous rate) of the midbrain dopmaine neurons; and for worse-than-expected outcomes (negative RPE), there is phasic decrease; no changes from spontaneous activity-rate is observed for same-as-expected (zero RPE) outcomes (Hart et al., 2014; Schultz et al., 1997; Tobler et al., 2005). Paralleling this, BOLD signal in human midbrain areas, such as the ventral striatum, ventral tegmental area, is also modulated by RPE (D'Ardenne et al., 2008; Pessiglione et al., 2006). Also, FRN may be indirectly associated with the midbrain dopamine activity. Multiple studies have suggested that the neural-generator of the FRN is located in the anterior cingulate cortex (ACC; Gehring & Willoughby, 2002; Hauser et al., 2014; Miltner et al., 1997; Van Veen et al., 2004); a site that receives midbrain dopaminergic input, through the mesocortical dopamine pathway (Rolls et al., 2008). Thus, the FRN may result from relaying of midbrain dopaminergic responses to the ACC region (for a review, see Walsh & Anderson, 2012).

If the FRN reflects a reinforcement-learning error signal, it may also support trial-to-trial learning. In reinforcement learning, larger (negative) RPE for errors are likely followed by response-adjustments in subsequent trials. Likewise, larger deflections of the FRN (for errors) may be followed by correct responses in subsequent trials. Supporting this idea, van der Helden et al. (2010) found that in a motor sequence-learning task, FRN amplitude was more negative when negative feedback outcomes (due to incorrect responses) were followed by correct responses, than when they were followed by incorrect responses. Another study by Cohen et al. (2007), who used a gambling task, found that for the high-rewarding stimulus, FRN amplitude became smaller with the course of learning, suggesting that participants learned to expect a reward more strongly for the high-rewarding stimulus. For the low-rewarding stimulus, FRN amplitude became larger with learning, suggesting that participants were not able to predict a reward strongly for the low-rewarding stimulus. Others have also reported larger FRNs for earlier- than later parts of the task (Walsh & Anderson, 2011b; Bellebaum & Daum, 2008), suggesting FRN-guided learning effects. Instead of immediate response-adjustments in the next trial, overall greater probability for subsequent response-adjustments for larger FRN, has also been noted by some studies (Yasuda, Sato, Miyawaki, Kumano, & Kuboki, 2004).

However, others have refuted the role of the FRN in indexing subsequent response-

adjustments (van de Vijver, Ridderinkhof, & Cohen, 2011; Luft, Takase, & Bhattacharya, 2014). For example, Chase et al. (2011) used a gambling (probabilistic learning) paradigm, which included multiple reversal-learning phases. In the reversal learning, the reward-contingencies for the high- and low-rewarding stimuli were reversed. Although the FRN amplitudes agreed with model-derived RPE-values for the task, they did not index subsequent response-adjustments in response to the reversal. Importantly, Chase et al. (2011) had instructed their participants, at the beginning of the experiment, that there will be sudden changes in reward-contingencies. Having this knowledge in the background may have recruited the FRN differently. Another study by Arbel et al. (2014) used a task where participants learned to associate novel-objects with non-words, through trial-and-error, and with feedback. The material were learned through repetitions (cycles). They found larger FRN for correct-feedback trials in cycle 1, when those were followed by correct responses in cycle 4, than when those were followed by incorrect responses in cycle 4. According to the FRN-RPE account, FRN amplitude for errors, when followed by correct responses, would be larger than when they are followed by incorrect responses. However, no such effect was found. Thus, the FRN may not support learning following the RPE function used in reinforcement learning theories.

Notably, the task used by Arbel et al. (2014) was substantially different from those used more commonly to study the characteristics of the FRN, such as, gambling tasks and time-estimation tasks. This could have also contributed to the findings reported above. Gambling tasks include a very small number of stimuli, each associated with a different reward-probability. On each trial, the participant bets on a stimulus for the better payout. Thus, the goal is to find a suitable response-strategy that helps maximize the total rewards. Accordingly, reward-predictions are used to drive the response-strategies. On the other hand, in time-estimation tasks, participants learn to be increasingly more accurate to respond within a specific time-interval, with feedback. Thus, here, learning reflects a conditioned motor-response, guided by the feedback signal.

Different from both gambling and time-estimation tasks, in the task used by Arbel et al. (2014), reward-prediction was used to guide response-adjustments based on the associations between novel-objects and non-words. The total number of stimuli was substantially higher ($N = 60$) than gambling tasks, which typically use 2 to 4 stimuli. Thus, memory for one novel-object and non-word association was subject to interference from the other novel-object

and non-word associations. This difference in the role of the reward-prediction, between gambling or time estimation tasks and learning tasks such as the one used by Arbel et al. (2014), could be important to understand the boundary conditions on the functions of the FRN. The theoretical RPE function only measures the difference between expected- and actual outcomes. However, the RPE-like function supported by the FRN may be subject to the number of learnable stimuli (see Collins & Frank, 2018). Others have also reported that the FRN amplitude becomes smaller when the information offered by the feedback is increased (Cockburn & Holroyd, 2018).

Importantly, the task used by Arbel et al. (2014) included only three cycles of learning for 60 pairs of novel objects and non-words. This may not have provided enough training for the associations, specifically for participants with slower learning rates. Moreover, Arbel et al. (2014) used performance in cycle 4 to evaluate the quality of learning, and compared successful and unsuccessful learning in terms of their FRN amplitudes in cycle 1. However, since RPE is trial-specific, the FRN-indexed RPE for cycle 1 may not be as relevant for the learning outcomes in cycle 4, as it would be for learning outcomes in cycle 2. Thus, Arbel et al. (2014) addressed how the very first feedback received in a trial-and-error learning situation has a long term effect on the behavioural adjustments.

In contrast, here we test if FRN can support trial-to-trial learning, where the evidence for learning is subsequent behavioural adjustment. We recently investigated characteristics of the FRN for a trial-and-error learning task, that also involved a large number of stimuli (words, $N = 48$), and reward feedback (see Chapter 4). Participants learned whether or not to choose each word depending on its inferred value (high or low). Through repetitions (16 cycles), most participants learned the stimulus-response rules considerably well. Thus, a surprise reversal, which reversed half of the stimulus-response relations at random, was designed to induce strong expectation-violations. Analysis of the feedback-locked ERPs for the surprise reversal cycle 17 elicited a negative-going deflection with similar latency as the FRN but with relatively more frontal than mid-frontal distribution of voltage across the scalp. Also, this FRN-like signal was not significantly larger for the switched- than the non-switched words. Follow-up analyses revealed that the FRN-like signal was influenced by an RPE-like function, but it was further modulated by task-variables that were not designed to manipulate the RPE. Further, re-learning of the stimulus-response associations, after the surprise reversal, was indexed by larger amplitude of the FRN-like signal, but only for one

subgroup of participants, who had not altered their previously learned responses, in the wake of the surprise reversal. Taken together, the FRN-like signal may have supported learning differently from the theoretical RPE function, and non-traditional tasks could help unravel this suggestion.

However, the interpretations above are from the traditional ERP effects of the FRN, which are limited to trial-averaged data only, and thus, do not provide direct insights for trial-to-trial learning, at the level of individual trials. For example, for the task described above, we found that the amplitude of the FRN-like signal, for the correct feedback-outcomes in cycle 1, was larger when those were followed by correct responses in cycle 2 than when followed by incorrect responses in cycle 2 (see Chapter 4). However, following traditional ERP methods, here, these amplitudes were averaged for trials that were correct in cycle 1 and incorrect in cycle 2, and then compared to the average for trials that were correct in cycle 1 and also correct in cycle 2. This difference in the average amplitudes of the FRN-like signal do not suggest that the distribution of its amplitudes for the two conditions mentioned above, were meaningfully different, when considering individual participants.

Here, to better estimate the role of the FRN-like signal and other relevant signals present during processing of feedback information, in trial-to-trial learning of the stimulus-specific response-rules for the task used in Chapter 4, we used predictive analysis, tailored to individual trials. We adapt the approach taken in Chapter 2 (Chakravarty et al., 2020) to assess the predictive value of brain activity during study trials of a verbal item-recognition task: we start with *a priori* univariate ERP measures that have been replicated numerous times, and follow that with data-driven, multivariate classifier analyses.

Predictions are meant to forecast the outcomes for future observations, whereas the planned-comparisons approach followed by traditional ERP analysis aims to explain current observations only; this is subject to overfitting, which is an error that is produced when a model fits to the noise in a sample dataset. An overfitted model is a perfect fit to the sample data it had operated on, but fails for other samples drawn from the same population, which have differently sampled random noise. Thus, the planned-comparisons approach runs the risk of overestimating the chosen neural measures by overfitting them; and the planned-comparisons approach can also underestimate the behavioural relevance of brain activity, by ignoring subtle multivariate patterns.

First, based on the knowledge from previous planned-comparisons-driven ERP effects of

the FRN (e.g., Cohen et al., 2007; van der Helden et al., 2010), we tested if the amplitude of the FRN-like signal classified between subsequently correct and incorrect responses. In other words, given the amplitude for a trial in cycle n–1, can we predict its response (correct or incorrect) in cycle n? As mentioned before, the FRN-RPE account would predict a larger (more negative) FRN amplitude for error trials that are followed by a correct response, than error trials that are followed by an incorrect response. If this relationship between the FRN amplitude and subsequent response accuracy is true, then we may be able to predict subsequent response accuracy for individual trials, based on its FRN amplitude from its previous trial. Thus, based on a classification rule that is derived from previous planned-comparisons analysis, we tested for predictions of subsequent response accuracy, for individual trials, and separately for each participant.

If the FRN-like signal found in this task does not support trial-to-trial learning following an RPE-like function, then the above classification-rule will not hold for individual trials. However, since the trial-and-error learning task was shaped by information retrieved from the feedback, it is likely that there exists other brain-activity signals during feedback processing that support trial-to-trial learning. Then, to identify those signals, or a combination of them, our next step was to follow exploratory, data-driven, multivariate pattern analysis methods. Thus, in this case, we did not pre-assign a rule for classifying subsequent responses into correct and incorrect trials. Instead, the relation between feedback-locked multivariate activity and subsequent response-accuracy was learned from the data itself, and separately for each participant. More importantly, the classifiers were trained and tested on different sets of trials, to check for overfitting.

To avoid forcing the result or selective reporting, we take the same approach to the classifier methods as in Chapter 2 (Chakravarty et al., 2020) and also in Chapter 3. Two, arguably the most simple, and linear classifiers were used: linear discriminant analysis (LDA; Fisher, 1936) and support vector machines (SVM; Cortes & Vapnik, 1995), both of which are well suited to learn linearly-separable multivariate patterns of brain-activity, representing different conditions, in this case, subsequent response-accuracy (correct or incorrect). Also, SVM is more robust against overfitting. Further, with LDA, it is straightforward to analyze the classifier-identified patterns (see Methods), and can be used to examine the relative importance of different signals for predicting subsequent response-accuracy in this task.

In sum, we evaluated the relevance of the FRN-like and other brain-activity signals

for learning the stimulus-specific response-rules in the task presented in Chapter 4, by using those signals to directly predict subsequent response accuracy for individual trials. Although the design of trial-and-error learning tasks suggests that information acquired from the feedback is crucial for learning, here we objectively measure the importance of brain-activity during feedback-processing, against predictive-benchmarks.

5.2 Methods

Participants We used data from the 58 participants reported in Chapter 4. All participants were native English speakers, had normal or corrected-to-normal vision and provided written informed consent in accordance with a University of Alberta ethics review board. Detailed description of the task and materials can be found in Chapter 4.

Behavioural materials and procedure Participants completed a "word choice task". A set of 48 words were used, half of which were chosen at random, at the beginning of the experiment, to be of high-value, the other half was set to be of low-value. Participants learned the value of each word through trial-and-error.

On each trial, they were presented (visually) with two items on either side of the computer screen (see Figure 4.1 in Chapter 4 on page 133). One of these items was always a word (common noun) and the other was always a nonsense string ('HHHHH'). They were instructed that there were two types of words - high- and low-value. For trials with high-value words, if they chose the word (pressed a button on the same side as the word) instead of the string 'HHHHH', they got a 10 point reward, indicated by an image of a pile of coins. Also, in this case, choosing the string 'HHHHH' led to 1 point, where the feedback was an image of one coin. For the trials with low-value words, the rule was opposite, i.e., choosing the string 'HHHHH' led to 10 points, and choosing the word led to 1 point. The contrast for the two feedback images, for 1 and 10 points rewards, was equated. The goal for the participant was to maximize the reward points, which, at the end of the session, were converted to a small monetary bonus (up to CAD 5) that they received in addition to course credit for participation.

All words were displayed (in random order) in each of 19 cycles. However, in cycle 17, half of the words switched their value, without warning. This meant that half of the previously

high-value words were randomly chosen to be of low-value and half of the previously low-value words were randomly chosen to be of high-value. The last two cycles (18 and 19) followed with no further changes to word-values

All responses were self-paced, but very fast responses (¡200 ms) were flagged and notified to the participant to prevent speeding through the task. There was a jittered interval of 500–700 ms between the response and the onset of the feedback as well as between the feedback offset and the onset of the next trial. Each feedback trial lasted 1500 ms.

EEG Recording EEG was recorded with high-density 256-channel Geodesic Sensor nets (Electrical Geodesics Inc., Eugene, OR), in an electrically shielded, sound-attenuated chamber. The raw signal was amplified at a gain of 1000. We used a sampling rate of 500 Hz for the recording and the impedance was kept below 50 $k\Omega$. The vertex electrode Cz was used as the reference. Preprocessing of the signal was done using the EEGLAB toolbox (`http://sccn.ucsd.edu/eeglab`; Delorme & Makeig, 2004). Preprocessing steps included bandpass filtering (0.5 Hz–30 Hz), average re-referencing and decomposition with independent component analysis (ICA) to identify artifactual activity such as, eye blinks, channel noise and muscle noise. Then, individual epochs relative to the onset of the feedback were extracted from the signal. Each epoch included a pre-stimulus onset window of 200 ms and a post-stimulus onset window of 1000 ms. Baseline, for each trial, was calculated by averaging the signal over the 200 ms pre-stimulus interval and was subtracted from all values in each particular trial. To detect epochs containing artifacts, an absolute voltage threshold of 200 μV was used. We also excluded trials based on a threshold of 25 μV point-to-point difference. On average, 1.2% trials were rejected per participant with these thresholds.

5.2.1 Predictions with the amplitude of the FRN-like signal

We tested if the amplitude of the FRN-like signal for a word in cycle n–1 predicted its response-accuracy in cycle n. Following a signal-detection theory approach (Green & Swets, 1966), previously also used in Chapter 2 (Chakravarty et al., 2020), first, we calculated the amplitudes (mean-voltage over the 200–350 ms time window, post feedback-onset) for individual trials. Then, these voltage measures were sorted by their magnitude. After that, we set a variable voltage-threshold, classifying trials above the threshold as subsequently-incorrect, and those below the threshold as subsequently-correct. This classification-rule was

based on the FRN-RPE account— feedback trials that are followed by correct responses are associated with a more negative FRN than trials that are followed by incorrect responses. Then, the true-positive and false-positive rates, corresponding to each voltage-threshold, were plotted against each other, to obtain the receiver operating characteristic (ROC) curve. The area under the curve (AUC) of the ROC was used to measure classification success. For AUC = 0.5, classification was at chance; for AUC = 1, classification was perfect.

5.2.2 Multivariate pattern analysis

All classification analysis followed the same rule for selecting the time-domain features of EEG activity (during feedback processing), as was also followed in Chapter 2 (Chakravarty et al., 2020) and in Chapter 3. First, to avoid any circularity in the classification logic, we pre-selected a set of 10 electrodes, roughly covering the scalp (see Figure 2.3 in Chapter 2 on page 43). Then, for each trial, the signal from each electrode was binned into 100 ms long time-bins. This produced a total of 100 features per trial— 10 electrodes and 10 time-bins. The classifiers were trained with 5-fold cross validation, the cross-validation folds were stratified, which means that the ratio of the number of trials, for the two classes (subsequent-correct and subsequent-incorrect), was kept constant. Further, the trial numbers for the two classes were balanced within each training set, using the Synthetic Minority Oversampling Technique (SMOTE; Chawla et al., 2002), in order to prevent the bias in classifier-training for the over-represented class; however, note that previous results from Chapter 2 (Chakravarty et al., 2020) suggested that oversampling by SMOTE did not alter the effects significantly from that without oversampling. To reduce chances of overfitting, both LDA and SVM were regularized, the regularization parameter was set at 0.5 for both. Classifier performance was measured by the area under the curve (AUC) of the receiver operating characteristic curve (ROC); AUC was averaged over the 5 test folds.

5.2.3 Analysis of classifier-identified patterns

With LDA, it is straightforward to look into the features of importance, as determined by the classifier. The coefficient of each feature in the model directly translates into the weight or importance of the feature relative to all other included features. For each participant, these weights were first averaged across the 5 training folds. Then, to compare across participants, weights were re-scaled to the $[0, 1]$ interval. Finally, average weights across participants were

used to look into the relative importance of the 10 time-bins as well as plotted on the scalp to look into the relative importance of the different electrodes/scalp regions, separately for each time-bin.

5.2.4 Finding the *steepest* cycle for predicting subsequent response-accuracy

We considered that predicting subsequent response-accuracy of cycle n, based on feedback-related brain-activity measures in the previous cycle n–1, would be more relevant for a pair of successive cycles that reflect a decent increase in accuracy. This increase in accuracy would suggest that the response-rules, for multiple trials, may have been acquired in cycle n, relative to their response-status in cycle n–1. Although, on average (across participants), accuracy for the early training cycles increased sharply (see Figure 4.2a on page 136), the amount of increase in accuracy for a given pair of successive cycles varied across participants (e.g., see Figure 4.4c on page 138). Thus, instead of choosing the same pair of successive cycles across all participants to test the predictions for subsequent response-accuracy, a better approach would be to choose the pair of successive cycles that show the most increase in accuracy, and are selected individually for each participant.

On the other hand, for testing the predictions, it is also necessary that the later cycle n contains both correct and incorrect trials. Accordingly, cycles with perfect accuracy cannot be used as the later cycle n, even if it showed the biggest increase in accuracy with respect to its previous cycle. Thus, taking both of the above considerations into account, we came up with the following rule to select the *steepest* cycle for each participant.

1. For each participant, first, we located the training cycles for which accuracy was greater than or equal to 75% but less than 100%. Then, among those cycles, we selected the one with the least accuracy. This means that we chose the cycle with accuracy closest to the 75% accuracy criterion. Across participants, the maximum cycle-accuracy obtained this way was 85%.

2. Next, we considered the accuracy of all cycles between (and including) cycle 1 and the cycle chosen above, and calculated the pairwise difference in their accuracy ($accuracy_n - accuracy_{n-1}$).

3. Then, we located the pair of successive cycles with the biggest, positive difference in accuracy.

Thus the *steepest* cycle (n) chosen for each participant included both correct and incorrect trials to test the predictions, and at the same time, reflected the biggest positive change in accuracy with respect to its previous cycle (n–1). However, the amount of change in cycle-accuracy for the *steepest* cycle with respect to its previous cycle varied across participants. For participants with a slower learning rate, the above change would have been small, and for those a faster learning rate, it would have been larger.

5.2.5 Separating trials based on previous feedback-outcomes

We also considered that when predicting subsequent response-accuracy, it would be a cleaner comparison to separate trials that received correct- and incorrect feedback-outcomes in the previous cycle. For error trials in the previous cycle, that are subsequently responded correctly, the FRN-RPE account would predict a bigger signal change, than when these are not responded correctly in the next cycle. On the other hand, for correctly-responded trials in the previous cycle, the interpretation is not straightforward. For example, if the previous (correct) response was a guess, then a subsequent correct or incorrect response could follow the same logic as above. However, if the previous response was not a guess, and was a learned response instead, then a subsequent correct response would be maintaining that learned information, whereas a subsequent incorrect response could be due to failure to remember; as discussed in the introduction, due to the large number of stimuli present in this task, it is possible to fail to remember the correct response for a word, even if it was learned before.

Moreover, not separating the trials based on previous feedback-outcome also introduces a potential confound to the classifier analysis. As we unpack later in the Discussion, classifiers, due to being data-driven techniques, can be influenced by circumstantial features. Here, the classifiers are tasked to predict response accuracy of cycle n, based on the feedback-locked multivariate brain activity features present in cycle n–1. However, the accuracy of the two cycles correlate with each other. Accordingly, the classifier could simply be predicting the accuracy for cycle n–1 itself.

Thus, for both the predictions with the FRN amplitudes and with the multivariate pattern analysis, we present the results with and without separating the trials based on previous

feedback-outcomes.

5.2.6 Statistical analysis

All analyses were done using MATLAB (2018b) and specific functions from the Statistics and Machine Learning Toolbox (Martinez et al., 2017). To test classifier success across participants, two-tailed t-tests (against chance) were used, significant effects were relative to $\alpha = 0.05$. The 95% confidence intervals of the mean were used as an estimate of the size of the prediction. Bayesian t-tests were also conducted, using a MATLAB function written by SamPenDu (2015). The Bayesian probability for the alternative- over the null hypothesis is indicated by the Bayes factor (BF_{10}); $BF_{10} = \frac{p(H_1)}{p(H_0)}$. For $BF_{10} > 10$, there is strong evidence for the alternate and for $BF_{10} < 0.1$, there is strong evidence for the null (Kass & Raftery, 1995). For $BF_{10} > 3$, there is moderate evidence for the alternative and for $BF_{10} < 0.3$, there is moderate evidence for the null. A pseudo-random number generator (Mersenne twister, seed = 0) was used for reproducibility of the classification results.

5.3 Results

Behaviour As reported previously in Chapter 4, participants performed at chance for cycle 1, which was expected (Figure 4.2). After that, performance increased sharply with increasing cycles and approached the maximum by cycle 16. Thus, participants were able to learn the stimulus-response associations of the 48 words within the 16 cycles of value learning. For the earlier training cycles, performance for high-value words tended to be greater than that for the low-value words (see Figure 4.2a on page 136). This is because participants preferred to choose the word more often than the string 'HHHHH' (word choice bias). Also, response times for the low-value words tended to be slower than that for the high-value words (see Figure 4.2b on page 136, considering correct responses only). This could be because low-value words required incongruent responses (press a button on the opposite side of the word).

Average performance for the last four training cycles (cycles 13 to 16, see Figure 4.3a on page 137) showed that a small number of participants ($N = 11$) did not learn the word-values considerably well (average performance accuracy less than 0.8). These were flagged as 'non-learners'. Also, analysis of performance in the reversal cycle 17 showed

that participants either responded on the basis of knowledge acquired in the training cycles ($N = 21$, conservative, see Figure 4.4a on page 138, or changed it in the wake of the reversal ($N = 26$, exploratory, see Figure 4.4b on page 138).

Analysis of the feedback-locked ERPs during the surprise reversal (cycle 17) showed that the amplitude of the FRN-like signal was not significantly more negative for the switched- than the non-switched words (see Chapter 4). However, since the exploratory participants had frequently altered their responses, we speculated that whenever a participant altered their previously-learned response for a trial, their reward prediction for that trial may have changed as well. Supporting this, we found that the amplitude of the FRN-like signal for cycle 17 was significantly more negative for switched and non-switched trials when the responses were not altered. Further, acquiring the new response-rules in cycle 18 was indexed by the amplitude of the FRN-like signal in cycle 17— its amplitude in cycle 17 for switched trials that were followed by correct responses in cycle 18 was more negative than those followed by incorrect responses. However, this response adjustment indexed by the FRN-like signal for cycle 18 was significant for the conservative- and not the exploratory strategy participants (see Chapter 4).

Notably, for the training cycles, there was no significant difference in learning rates between the groups. In our classification analysis, we present results for all participants, followed by the breakdown of participants into non-learners and the two strategies.

5.3.1 Feedback-locked ERPs

Detailed analysis of the feedback-locked ERPs for the surprise reversal cycle 17 can be found in Chapter 4. Here, we looked into the feedback-locked ERPs during the early training cycles, to get a sense of what to expect from the classification analysis that follows. Non-learners were not included in the ERP analysis.

We considered the *steepest* cycle (the steepest increase in accuracy from one cycle to the next), which was chosen individually for each participant, following the rules described in the Methods. Then feedback-locked ERPs from its previous cycle were obtained and plotted separately correct and incorrect trials in the *steepest* cycle, as well as, for correct and incorrect trials in the previous cycle (see Figure 5.1). For high-value words that were correctly responded to in the previous cycle, amplitudes of the FRN-like signal for subsequently correct trials appeared to be more negative than subsequently incorrect trials (Figure 5.1a). A paired

t-test between these two conditions was significant, $t(37) = 2.48,\ p < 0.05,\ BF_{10} = 2.55$, but the Bayes Factor ($BF_{10} < 3$) was inconclusive. However, for the other conditions, the difference in the amplitudes of the FRN-like signal for subsequent correct and incorrect responses was smaller (Figure 5.1b–d). Thus, to compare across all conditions, we conducted a 2×2×2×2 repeated-measures ANOVA for the amplitudes of the FRN-like signal, with the within-subject factors Value (high and low), Accuracy in the previous cycle (correct and incorrect), Accuracy in the subsequent cycle (correct and incorrect), and the between-subject factor Strategy (conservative and exploratory). However, this model revealed no significant main or interaction effect for Accuracy in the subsequent cycle. Also, note that the degrees of freedom of the model was reduced because multiple participants did not make any incorrect response for previously correct trials, and thus, they were excluded from the analysis.

Overall, for the high-value and previously-correct trials the ERP analysis showed a significant effect in favour of the suggestion that the FRN-like signal in this task indexed trial-to-trial learning. However, this was not true for the other relevant conditions.

5.3.2 Predicting subsequent response-accuracy with the amplitude of the FRN-like signal

Considering the *steepest* cycle, which was individually chosen for each participant (see Methods), the predictions based on the amplitudes of the FRN-like signal are presented in Figure 5.2. When all trials from the previous cycle were considered, this amplitude did not predict subsequent response-accuracy significantly better than chance (0.5); $t(50) = 1.48,\ p = 0.15$, 95% $CI = [0.49\ \ 0.55],\ BF_{10} = 0.42$, but the Bayes Factor ($0.1 < BF_{10} < 3$) was inconclusive. When only incorrect trials from the previous cycle were considered, the amplitude of the FRN-like signal also did not predict subsequent response-accuracy significantly better than chance; $t(50) = 0.70,\ p = 0.49$, 95% $CI = [0.47\ \ 0.56],\ BF_{10} = 0.19$, also with inconclusive BF_{10}. However, when only correct trials from the previous cycle were considered, there was a significant effect: $t(48) = 2.92,\ p < 0.01,\ 95\%\ CI = [0.52\ \ 0.61],\ BF_{10} = 6.51$; the Bayes Factor ($3 < BF_{10} < 10$) suggested moderate evidence in favour of the effect. The 95% confidence interval suggested that the size of the prediction was modest. Moreover, breaking down the results into the two strategies showed that the effect was significant for the exploratory- ($t(23) = 2.43,\ p < 0.05,\ 95\%CI = [0.51\ \ 0.62]$) and not the conservative ($t(20) = 1.30,\ p = 0.21,\ 95\%CI = [0.47\ \ 0.63]$) strategy participants (Figure 5.2). Given

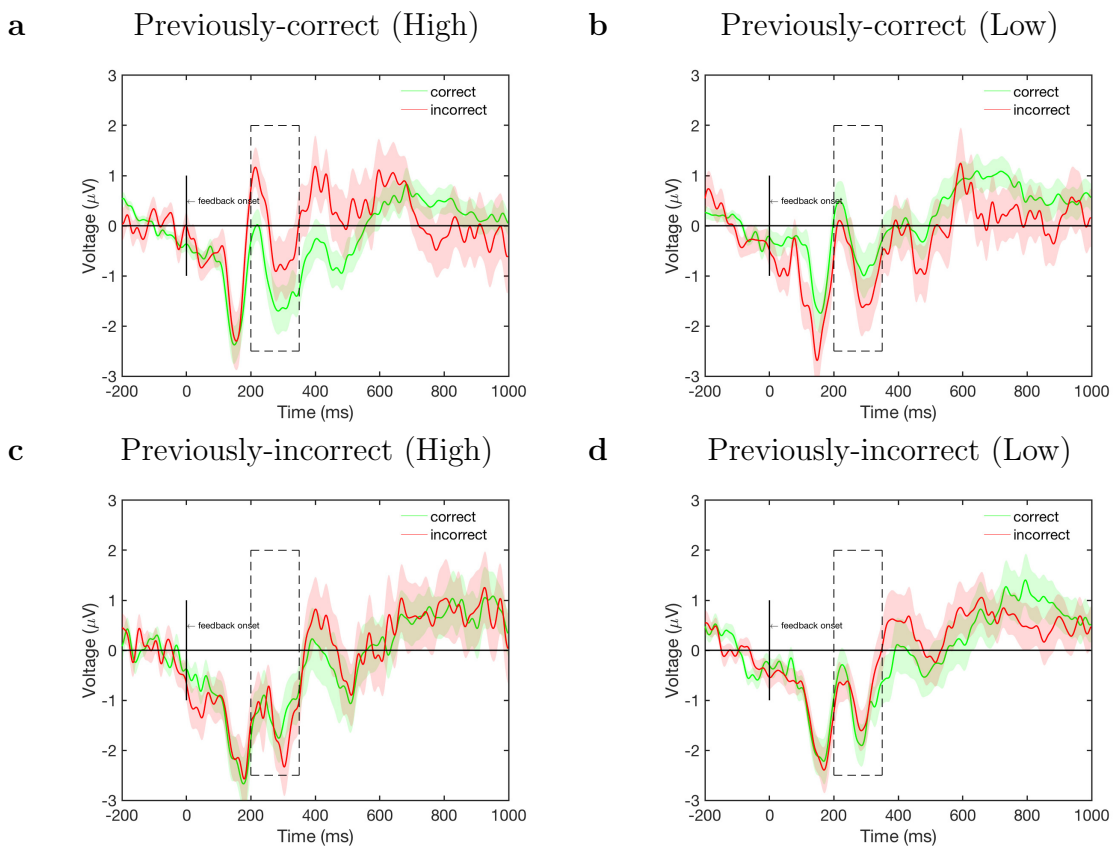

Figure 5.1: Grand averaged ERPs for learners ($N = 47$) during feedback presentation, and as functions of response accuracy in the *steepest* cycle, chosen individually for each participant (see Methods). ERPs are shown separately for high- and low-value (left and right panels), for correct and incorrect responses in the subsequent cycle while keeping response accuracy restricted to correct (upper panels) and incorrect (lower panels) trials for the preceding cycle. Shaded error bars are standard errors of the mean. All ERPs are plotted for the fronto-central electrode FCz.

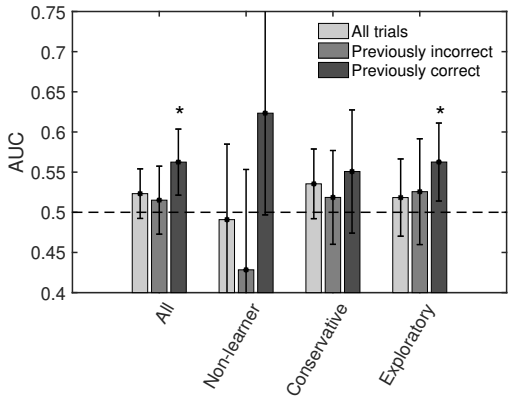

Figure 5.2: Classification of subsequent response-accuracy, based on the amplitudes of the FRN-like signal, and for the *steepest* cycle, chosen individually for each participant (see Methods). Classifications are presented separately for when all trials were considered in the previous cycle, when only previously-incorrect trials were considered, and when only previously-correct trials were considered. Results are also shown separately for all participants (N=58), non-learners (N=11), conservative- (N=21) and exploratory (N=26) strategy participants. Dashed line presents chance. Error bars are 95% confidence intervals. Significant effects are marked with *.

that the means were similar, this could also be due to slightly more participants in the exploratory- than conservative strategy group.

Thus, the amplitude of the FRN-like signal supported trial-to-trial learning for the correct responses only. As discussed before, for previously-correct trials, if the response was a guess, then a greater signal change in the FRN-like signal for later correct responses could agree with the FRN-RPE account. However, considering this result, along with the non-significant effect for the previously-incorrect trials, it is possible that the FRN-like signal in this task did not follow an RPE-like function in order to support trial-to-trial learning of the correct responses. Notably, this result was consistent with the conclusions of Chapter 4. Instead, the FRN-like signal may have indexed maintenance of learned responses across successive cycles, with a greater signal change. Notably,

5.3.3 Predicting subsequent response-accuracy with multivariate pattern analysis of brain-activity during feedback-processing

Next, we tested if multivariate pattern analysis of brain activity during feedback-processing predicted subsequent response-accuracy. Following the same approach as above, responses were considered for the *steepest* cycle, chosen individually for each participant. The re-

sults showed that when all trials were considered from the previous cycle, both LDA and SVM achieved significant success in predicting subsequent response-accuracy (see Figures 5.3 and 5.4); LDA: $t(50) = 2.54,\ p < 0.05,\ 95\%\ CI = [0.51\ \ 0.58],\ BF_{10} = 2.74$, SVM: $t(50) = 2.60,\ p < 0.05,\ 95\%\ CI = [0.51\ \ 0.59],\ BF_{10} = 3.17$, but the BF_{10} provided moderate support for the SVM effect only. Also, in this case, for both LDA and SVM, the effects were more significant for the exploratory- than the conservative strategy participants (Figures 5.3 and 5.4), which, once again, could be due to more participants in the exploratory strategy group. However, as mentioned before, the classifier success in this case, could be due to the classifiers predicting response accuracy for the previous cycle itself.

When only incorrect trials were considered from the previous cycle, both LDA and SVM failed to produce significant effects; LDA: $t(46) = 0.25,\ p = 0.81,\ 95\%\ CI = [0.46\ \ 0.56]$, $BF_{10} = 0.16$, SVM: $t(46) = 0.89,\ p = 0.38,\ 95\%\ CI = [0.47\ \ 0.58],\ BF_{10} = 0.23$; BF_{10} was inconclusive in both cases. Interestingly, in this case, we found a positive correlation between LDA performance and asymptotic accuracy of the participants (average accuracy over the last four training cycles); $r(47) = 0.31,\ p < 0.05$; there was also a trend-effect for SVM, $r(47) = 0.24,\ p = 0.09$. The correlation plots (see Figure 5.5) showed that this result was mainly due to the non-learners, for whom, the classifiers performed worse.

Finally, when only correct trials were considered from the previous cycle, both LDA and SVM once again failed to produce significant effects; LDA: $t(29) = 0.68,\ p = 0.50,\ 95\%\ CI = [0.46\ \ 0.59],\ BF_{10} = 0.24$, SVM: $t(29) = -0.16,\ p = 0.87,\ 95\%\ CI = [0.42\ \ 0.57],\ BF_{10} = 0.20$; BF_{10} was inconclusive in both cases. However, in this case, the degrees of freedom were substantially lower than all other analyses reported above.

Overall, multivariate pattern analysis of brain activity during feedback-processing did not succeed in predicting subsequent response accuracy. However, this set of analyses was subject to lower number of trials available for training the classifiers. Also, the degrees of freedom of the t-tests suggested that especially for the case of previously-correct trials, many participants did not have enough trials in order to conduct the classifier analysis with 5-fold cross-validation. Thus, the failure of the multivariate pattern analysis could be due to lack of power.

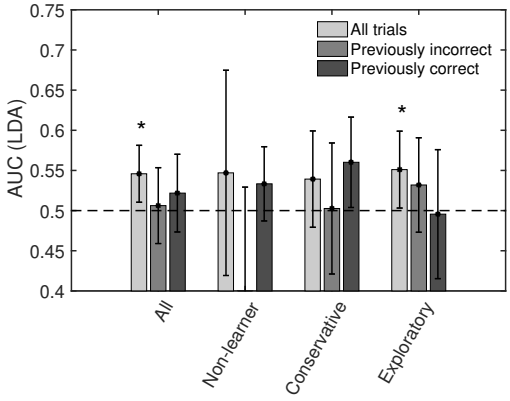

Figure 5.3: Classification of subsequent response-accuracy with LDA, based on multivariate pattern analysis of brain activity during feedback-processing, and for the *steepest* cycle, which was chosen individually for each participant (see Methods). Classifications are presented separately for when all trials were considered in the previous cycle, when only previously-incorrect were considered, and when only previously-correct trials were considered. Results are also shown separately for all participants, non-learners, conservative- and exploratory strategy participants. Dashed line presents chance. Error bars are 95% confidence intervals. Significant effects are marked with *.

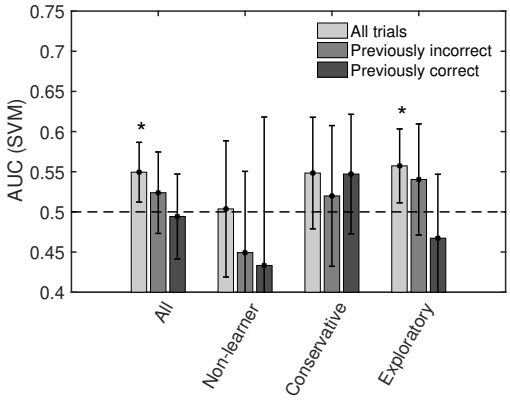

Figure 5.4: Classification of subsequent response-accuracy with SVM, based on multivariate pattern analysis of brain activity during feedback-processing, and for the *steepest* cycle, which was chosen individually for each participant (see Methods). Classifications are presented separately for when all trials were considered in the previous cycle, when only previously-incorrect were considered, and when only previously-correct trials were considered. Results are also shown separately for all participants, non-learners, conservative- and exploratory strategy participants. Dashed line presents chance. Error bars are 95% confidence intervals. Significant effects are marked with *.

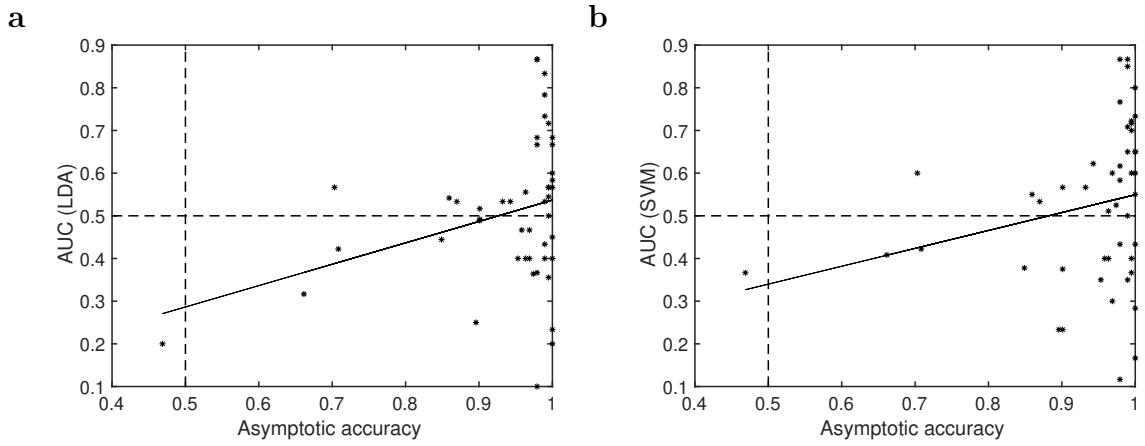

Figure 5.5: Correlation between classifier performance and asymptotic accuracy of the participants (average accuracy over the last four training cycles), separately for LDA (a) and SVM (b). The classification was for subsequent response-accuracy, restricted to previously incorrect trials (the middle bars for each group in Figures 5.4 and 5.3). Dashed lines present chance performance. Solid lines are the regression lines.

5.3.4 Predicting word-value with multivariate pattern analysis of brain activity during feedback processing

Lastly, we wondered if multivariate pattern analysis of brain-activity during feedback-processing predicted word-value (high or low), suggesting that participants were thinking back about the trial, which facilitated learning. Since classifying value was not confounded by the dependence of accuracy on cycle, to better train the classifiers with more trials, we pooled trials from cycles 2 to 5. However, due to the word-choice bias in the early training cycles, word-value was related to accuracy; high-value words were responded correctly more frequently than the low-value words. Thus, the classifier could pick up on the two different reward images used for the correct and incorrect feedback-outcomes, and predict word-value using this difference. To avoid this, we conducted the classifier analysis separately for when the feedback-outcome was correct or incorrect (see Figure 5.6a–b).

For correct feedback-outcomes, both LDA and SVM achieved significant success in predicting word-value (LDA: $t(57) = 3.37,\ p < 0.005,\ 95\%\ CI = [0.51\ \ 0.55],\ BF_{10} = 20.86$; SVM: $t(57) = 2.57,\ p < 0.05,\ 95\%\ CI = [0.51\ \ 0.54],\ BF_{10} = 2.90$), though only the LDA effect was strongly supported by the Bayes Factor ($BF_{10} > 10$). Both LDA and SVM also succeeded in predicting word-value when the feedback outcome was incorrect (LDA: $t(56) = 3.05,\ p < 0.005,\ 95\%\ CI = [0.51\ \ 0.57],\ BF_{10} = 8.97$; SVM: $t(56) = 2.17,\ p < 0.05$,

95% $CI = [0.50 \;\; 0.56]$, $BF_{10} = 1.25$); with BF_{10} indicating moderate support for the LDA effect only. Interestingly, the LDA weights for predicting word-value when the feedback outcome was incorrect showed a clear peak for the 5^{th} and 6^{th} time-bins, suggesting greater influence of the feedback-related signal in this time-window (see Figure 5.6c–d). On the other hand, the LDA weights for predicting word-value when feedback outcome was correct did not show a similar peak, and instead, they were overall higher for the earlier- than the later time-bins. However, the scalp distribution of the LDA weights for the 500–600 ms time-bin, as well as for an earlier 300–400 ms time-bin, looked very similar for the correct and incorrect feedback-outcome conditions (see Figure 5.6e–h). Thus, it is possible that there were same underlying signals, but those were weighted differently, and thus probably also recruited differently in the brain, when discriminating between high- and low-value words separately for correct- and incorrect feedback-outcomes. Taken together, these results suggest that brain-activity during feedback-processing reflected thinking back/processing of the word-value information, which may be implemented differently in the brain when the feedback-outcome was correct, and when it was incorrect.

5.4 Discussion

Our goal was to investigate brain activity while participants processed feedback information which supported trial-to-trial learning. We found that the amplitude of the FRN-like signal found in this task achieved significant success in predicting trial-to-trial learning when the previous trial was responded correctly. However, this amplitude did not predict trial-to-trial learning when an error was made in the previous trial or when both previously correct and incorrect trials were considered. Multivariate pattern analysis of brain activity during feedback processing failed to predict trial-to-trial learning for all three situations mentioned above. However, multivariate pattern analysis may have failed due to smaller number of trials available for training the classifiers, especially when considering previously-correct trials only. Interestingly, classifier performance increased with the performance of the participant; specifically, for participants who failed to learn the stimulus-specific response-rules even after 16 cycles of training, the classifiers also performed worse. Additionally, multivariate pattern analysis of brain activity during feedback processing achieved a small but significant success in predicting the inferred value of the word (high or low); and interestingly, the classifier-

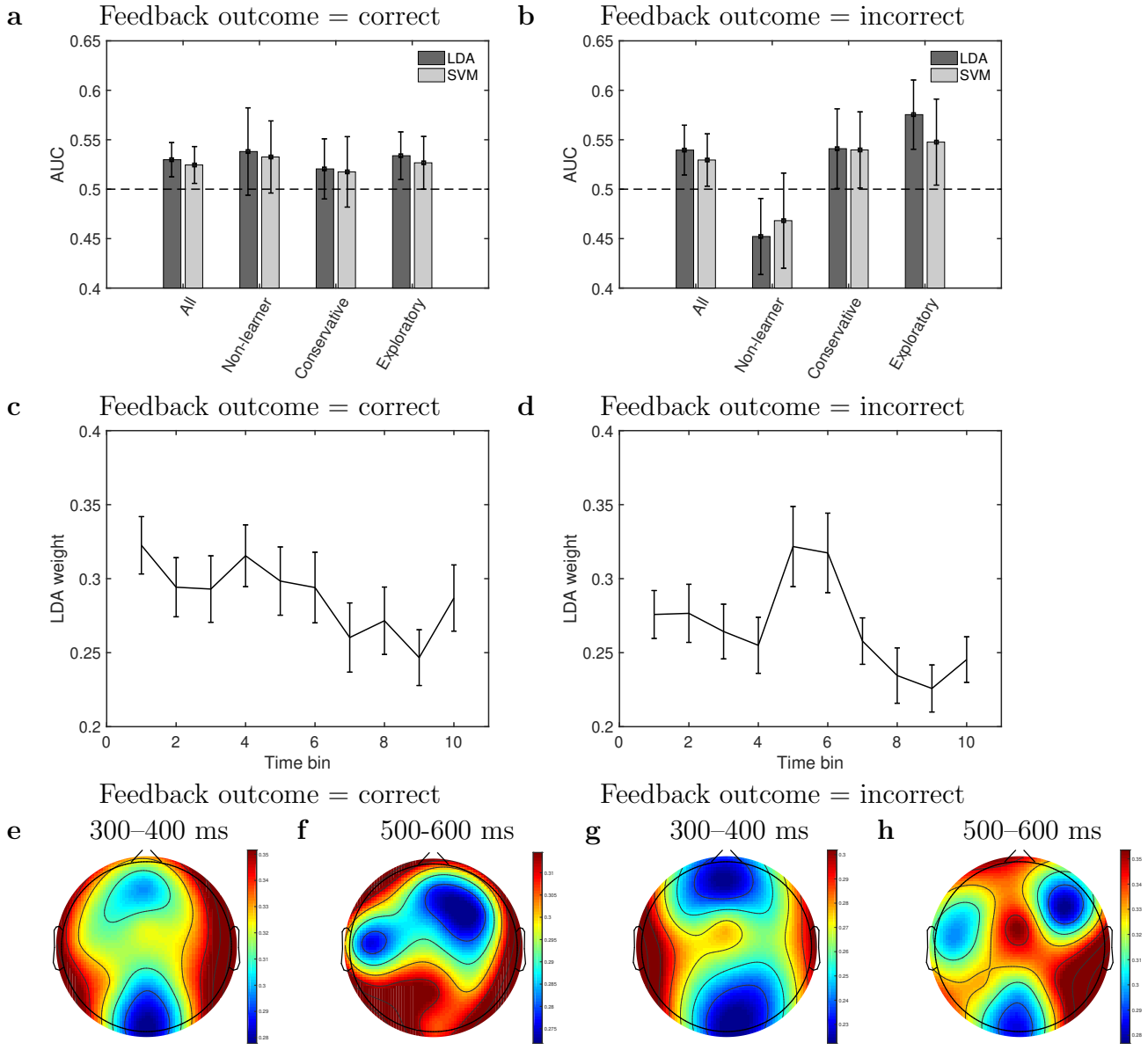

Figure 5.6: a–b. Classification of word-value (high/low), for trials pooled from cycles 2 to 5, based on multivariate pattern analysis of the feedback-related activity, for the same trials, and separately for when the feedback outcome was correct (a), and incorrect (b). Results are shown separately for all participants, non-learners, conservative and exploratory strategy participants. Dashed line presents chance. Error bars are 95% confidence intervals. c–d. LDA weights, averaged across all included electrodes, and across all participants with LDA AUC> 0.5, for classification of word-value, separately for correct- (c) and incorrect (d) feedback-outcomes. Error bars are standard errors. c–d. Scalp-distribution of the LDA weights, for the 300–400 ms, and the 500–600 ms time intervals. Color reflects weights, color axes are range-scaled.

identified pattern for discriminating between high- and low-value words was different for when the feedback-outcome was correct or incorrect. Together, our results offered novel insights for understanding brain activity behind trial-and-error learning, as we discuss below.

Trial-and-error learning presents a situation where information acquired from the feedback is crucial to do better in the task. Thus, it is natural to ask what features of brain activity may support learning from the feedback. Previous research suggests that the FRN may be a candidate. Specifically, the FRN-RPE account (Holroyd & Coles, 2002) suggests that the FRN acts like an reinforcement learning error signal in the brain. If this is true, then, similar to how the RPE function has been able to successfully explain reinforcement learning, the FRN could also support human trial-and-error learning. However, previous studies looking into this question have found mixed results. One possibility is that the FRN scales with an RPE function to support learning only for more traditional reinforcement learning tasks used in this type of research, such as gambling tasks or time estimation tasks. On the other hand, for tasks where reward prediction is not the central goal, but instead it guides stimulus-specific learning, such as the current task, the FRN-like signal may not 1) index a generic RPE computation (see Chapter 4) or 2) support learning following an RPE-like function. Notably, a study by Chase et al. (2011), who used a gambling task, also found that the FRN did not support trial-to-trial learning following an RPE function. Thus, functions of the FRN in human trial-and-error learning may be different from a theoretical RPE function.

Going beyond traditional ERP effects, here, we showed that amplitude of the FRN-like signal *predicted* learning from correct feedback-outcomes, but not from incorrect feedback-outcomes. These results contrast with what would be predicted by the FRN-RPE account—that for a trial, there is a greater probability for a change in response when it was previously responded incorrectly. The traditional ERP analysis of the FRN-like signal found in this task also showed more negative amplitude for correct trials when they were followed by correct trials in the next cycle, than when they were followed by incorrect trials in the next cycle, but only for high-value words and the Bayes Factor was inconclusive. Two other studies which used a paradigm close to ours, also found (with traditional ERP analysis) that FRN indexed learning for correctly-responded trials only (Arbel et al., 2013, 2014).

As discussed briefly in the Methods and Results, a correct trial could have been either a guess (for example, all correct trials in cycle 1 were guesses) or a learned response. A guess

would have produced a non-zero (positive) RPE, and accordingly, a larger FRN amplitude will be more likely if the guessed-correct response was followed by another correct response, than if it was followed by an incorrect response. This agrees with the FRN-RPE account. However, if this is true, then a larger FRN will also be more likely when an incorrect (which has to be a guess) response was by followed a correct response, than if it was followed by another incorrect response. However, this was not the case. Alternatively, a correct response could be a learned response, in which case, it would produce a zero RPE, and thus, a smaller FRN, and there is no response adjustment required in the subsequent trial. However, our results suggested that there was a greater signal-change for the FRN when a correct response was followed by another correct response, than when a correct response was followed by an incorrect response. Also, consider the interpretation for the situation where a correct response is followed by an incorrect response. Unlike gambling tasks, here, rewards were deterministic, and thus, once the response-rule for a word was learned, there was no need to bet on it in subsequent trials. Thus, only possible reason behind the situation where a correct response was followed by an incorrect response is that it was simply not remembered. This is possible, for there were 48 trials per cycle, producing interference in memory. Thus, a greater signal change in the FRN-like signal for correct responses that were followed by correct responses, than for correct responses that were followed by errors, could suggest that the FRN-like signal supports maintaining of correct or learned responses from one trial to the next. This contrasts with the RPE function, which supports response adjustments after making errors.

Although the FRN-like signal did not predict learning for previously-incorrect trials, there could be other signals present during feedback processing that do so. However, our multivariate pattern analysis of brain activity during feedback processing could not predict learning either. Classifiers require a large number of trials for better training. Accordingly, our initial plan was to pool trials from multiple cycles. However, classifiers are also sensitive to circumstantial information, in this case, the dependence between response-accuracy and training cycles. As the number of cycles increased, response-accuracy increased. Thus, a classifier designed to predict response-accuracy should also be able to make predictions about the cycle, albeit not perfectly. However, with trials pooled across cycles, the classifier could be doing the reverse. In other words, the classifier could be predicting response-accuracy based on cycle alone. Thus, trials were not pooled across multiple cycles. However, even for

predicting the response-accuracy for one cycle, based on the multivariate feedback-related activity from the previous cycle, the classifiers could have predicted the response-accuracy of the previous cycle itself, and produced a significant success, as we saw in the results. Thus, restricting to previously-correct or previously-incorrect trials only was a much cleaner comparison. However this reduced the number of training trials drastically, producing underpowered, non-significant effects. Importantly, predictions with the amplitudes of the FRN-like signal only were not subject to these confounds; in that case, all the trials were treated as test trials, and the classification-rule was obtained from previous ERP effects of the FRN, in accordance to the RPE account.

Notably, cognitive processes present during the study or encoding of information are thought to be important predictors of the variability in later memory success. In the *subsequent memory effect* approach (Sanquist et al., 1980), researchers have examined brain activity during the study phase of memory tasks to identify signals that index later memory-success at test. The subsequent memory effect approach is referred to as identifying brain activity that is predictive of memory, because it evaluates memory outcomes as functions of brain-activity signals that are operational at an earlier time than when the actual test happens (for a review, see Paller & Wagner, 2002). However, very few studies have looked into the predictability of the brain-activity signals identified with the subsequent memory effect framework using actual predictive tests (Chakravarty et al., 2020; Fukuda & Woodman, 2015; Noh et al., 2014; Sun et al., 2016; Watanabe et al., 2011). In a previous study (Chapter 2; Chakravarty et al., 2020), we found that the EEG (time-domain) features present during the study-phase of an item-recognition task were indeed predictive of later memory-success (at test), but the size of prediction, for both individual ERP-measures at study (late-positive component and slow-wave), as well as for multivariate pattern analysis of study-related activity, was small, though the classifier performance increased with participant's performance, and was meaningfully large for participants who performed well.

The subsequent memory effect framework applies to many learning situations, including the current task, where response-accuracy may be explained as functions of brain-activity signals present during feedback-processing of the previous trial. However, the subsequent memory effect has been more commonly studied in what can be referred to as 'one-shot' learning— participants study unique lists of items (such as, words) followed by memory-tests, such as old/new recognition, recall etc., and thus, the tests typically evaluate memory

outcomes after single exposure to the items only. Further, learning in those memory tasks is based on instruction only, there is no clear, immediate incentive for the participant to learn the lists, other than their general motivation to do well in the task. In contrast, here, the same set of 48 words were repeated across all cycles, and also, there were rewards, both of which could have made the task more engaging, producing more task-relevant brain activity. In partial support of the above suggestion, the size of prediction with the amplitude of the FRN-like signal (for previously-correct trials), as estimated by the 95% confidence intervals of the average AUC, was greater ([0.52 0.61]) than what we found with the amplitudes of the late-positive component ([0.51 0.54]) or the slow wave ([0.51 0.54]), when predicting subsequent memory success in Chapter 2 (Chakravarty et al., 2020). Note that the number of participants were comparable across the two studies. However, in Chapter 2 (Chakravarty et al., 2020), the number of trials included (for each participant) into the predictive analysis was substantially higher. Taken together, the current investigations draw an interesting parallel with the subsequent memory effect investigations, and it may be possible to find even more predictive success with multivariate pattern analysis of brain activity (during feedback processing) for the trial-and-error learning task by figuring out a better way to employ those methods.

In our task, learning entailed gaining knowledge about the value of a word (high or low). Accordingly, we asked if the feedback-related brain activity predicted the value of the preceding word stimulus. We found significant success with the multivariate classifiers in predicting word-value, but the size of the prediction was small. We conducted the analysis separately for correct- and incorrect feedback-outcomes, because the two reward-images used for correct- and incorrect feedback-outcomes were different, and as mentioned before, classifiers can be sensitive to such information— due to the word-choice bias in the earlier training cycles, correct (the 10 points reward) responses were more common for high- than low-value words, and thus, the classifier could be predicting word-value by learning about the difference due to the two reward images. Analyzing the classifiers separately for correct- and incorrect feedback-outcomes helped avoid this confound, and also produced an interesting result. Although classification success for both the correct- and incorrect feedback-outcome conditions were comparable, the LDA-weights suggested different underlying patterns. Notably, when the feedback-outcome was correct, it meant that the preceding response was either a congruent action for the high-value words, or an incongruent action

for the low-value words. For incorrect feedback-outcomes, this was opposite. Thus, the difference between the classifier-identified patterns could suggest how the participant learned to make congruent and incongruent responses to high- and low-value words, respectively. Taken together, learning from correct- and incorrect feedback-outcomes involved different signals, and potentially different cognitive processes underlying those signals, even though correct and incorrect feedback-outcomes were equally useful to learn the stimulus-specific response-rules.

To our knowledge, no other study has investigated brain-activity signals for trial-and-error-learning with predictive approaches, though research in the field of brain-computer interfaces (BCI) has focused on error-related signals such as the FRN to improve the performance of the BCI equipment (Mousavi & de Sa, 2019; Chavarriaga, Sobolewski, & Millán, 2014); this is a different goal. However, a recent study by Williams, Hassall, Lindenbach, and Krigolson (2019) may be relevant. Here, participants learned to associate Tamil/Manipuri symbols to English words through trial-and-error. A computational (reinforcement learning) model was also fitted to do the same task, in order to compare between experimental data and model behaviour. Their computation of the FRN amplitude was a little different from the current study, for their analysis of the FRN was based on an alternate theory suggested by some researchers previously. Specifically, some have suggested that the FRN may be the result of a superposition of two signals: a negative-going deflection with latency close to 200 ms, also known as the N2, which is followed by a positive-going deflection, also known as the reward-related positivity. The N2 is thought to index unexpected outcomes whereas the reward-related positivity indexes positive outcomes (Holroyd et al., 2008). The superposition of the two signals with opposite polarity gives rise to a more negative FRN for negative than positive outcomes (for a review, see, Proudfit, 2015). Williams et al. (2019) analyzed the reward-related positivity, and found that its amplitude decreased with learning. Also, the reward-related positivity was negatively correlated with accuracy, and positively correlated with response times, for both participant- and the model data. Together, the RPE effect indexed by the reward-related positivity, a signal that may underlie the FRN, was strongly correlated with learning. However, their study did not test for predictions beyond the correlations, and thus, as with any descriptive analysis, this effect could be subject to overfitting. Correlations between variables do not readily suggest that one can be predicted from another, unless the relation is also evaluated for out-of-sample data, which contains

differently sampled random noise.

Another study by Arbel and Wu (2016) is also relevant. Their study followed a similar paradigm as that of Arbel et al. (2014), participants learned to associate novel objects with non-words through feedback and with repetitions or cycles. Using correlations, Arbel and Wu (2016) also found that the FRN amplitude decreased with learning. However, a gradual decrease in an EEG signal across a session may also be due to other reasons, such as, electcrodes drying out, etc., and may not be due to learning only. Additionally, with logistic regressions, Arbel and Wu (2016) found that while a larger FRN for the positive-feedback in cycle 1 predicted learning-success at a later test in case of the negative-feedback, a smaller FRN in cycle 1 predicted learning-success at the later test. Thus, for the negative feedback, the FRN-indexed learning effect followed an opposite rule to that for the positive feedback. However, this was not a classifier analysis; the logistic regression did not use separate trials for training and testing the model, and thus, could have been subject to overfitting as well.

In light of the findings Arbel and Wu (2016), who used a paradigm that was closer to the current study than a gambling- or time estimation task, one possibility is that in our study, the amplitude of the FRN-like signal for the previously-incorrect trials also indexed trial-to-trial learning following a reverse classification-rule than that for previously-correct trials. However, in that case, the ROC analysis of the amplitudes would have produced AUC values significantly below chance, but this was not the case. Thus, unlike Arbel and Wu (2016), in this study, the trial-to-trial learning effect indexed by the FRN-like signal was not orthogonal for previously-correct and incorrect trials.

Overall, our results added novel insights to human trial-and-error learning, and challenged the equivalence between an FRN-like signal and RPE (also see Chapter 4). The multivariate classifiers could not be employed efficiently, in order to make the same predictions as those tested with the amplitudes of the FRN-like signal; this is a direction to be pursued in detail in the future. However, when they were not under-powered, the multivariate classifiers were successful in predicting word-value.

Notably, the overall size of prediction, both in case of amplitudes of the FRN-like signal and the multivariate classifiers, were modest, though classifier performance depended on participants' performance. Thus, it is possible that for better-performing participants, brain activity was more task-relevant, which produced a higher signal-to-noise ratio (SNR), and the classifiers were able to leverage it. Also, the time-domain features of EEG are more

likely to be influenced by the trial-to-trial variability in the latency of the EEG signals, than for example, the spectral-domain features of EEG, though some suggest that the latency of the FRN is relatively more stable across trials, and thus, it takes smaller number of trials to obtain a significant FRN effect, following traditional ERP methods (Marco-Pallares et al., 2011). In general, when conducting classifier analysis for EEG data, researchers have more commonly opted for the spectral domain features (for example, see Weidemann et al., 2019; Noh et al., 2014, who used EEG classification to predict different memory outcomes). However, even in those cases, the prediction sizes were never near perfect. Although previous studies have not very commonly discussed the interpretations of their results based on the overall size of prediction, this is an important point that needs further consideration, specifically when the size of prediction is modest. It is possible that on average a small size of prediction is what we can expect from classification of EEG, and the chance of success for the classifier methods needs to be viewed in terms of the task-relevance (or "task-resolution") of recorded brain-activity. Further, it may be possible to evaluate this task-resolution with the participants' performance (see also Chapter 2 Chakravarty et al., 2020 and Chapter 3, where we found similar suggestions). For the investigations pursued here, it is also possible that cognitive processes present during feedback-processing only explain part of the variability in trial-to-trial learning. Other sources of variability may include cognitive processes present during processing of the stimulus, making responses, etc.; these would also need to be explored in detail in the future.

In sum, we found that an FRN-like signal might be able to support trial-to-trial learning. However unlike RPE, which supports response adjustments after making an error, the FRN-like signal may support maintaining of correct or learned responses, from one trial to the next. This also calls for a reconsideration of the ways in which reinforcement learning in thought to inform human trial-and-error learning.

Chapter 6

General Discussion and Conclusion

In this thesis, I investigated brain activity underlying learning success, with predictive approaches. Using single-trial estimates of the time-domain features of electroencephalographic (EEG) recordings, I showed that it is possible to make predictions about learning-outcomes at the level of individual trials, and also to gain insights about behaviourally-relevant brain-activity signals. In this final Chapter, I discuss the general implications for the findings and questions for the future. I also note the significance of this work in using predictive analysis as the main approach, along with its limitations and future directions.

6.1 General Implications

The central idea behind the investigations pursued here was that variability in learning outcomes can be explained by cognitive processes preceding them in time, and thus, predicted from brain activity that likely captures those cognitive processes. The studies first evaluated the predictive value of previously known event-related potential (ERP) measures related to the behavioural outcomes, followed by more data-driven analysis of the multivariate patterns. Figures 6.1 and 6.2 present visual summaries of the main investigations with the two paradigms. Here, I discuss how each of these investigations adds to the existing knowledge about behaviourally-relevant brain activity.

6.1.1 Predictive value of univariate ERP measures derived from prior studies

Late positive component, slow wave, and the subsequent memory effect The two ERP signals during the study phase of the item recognition task, late positive component

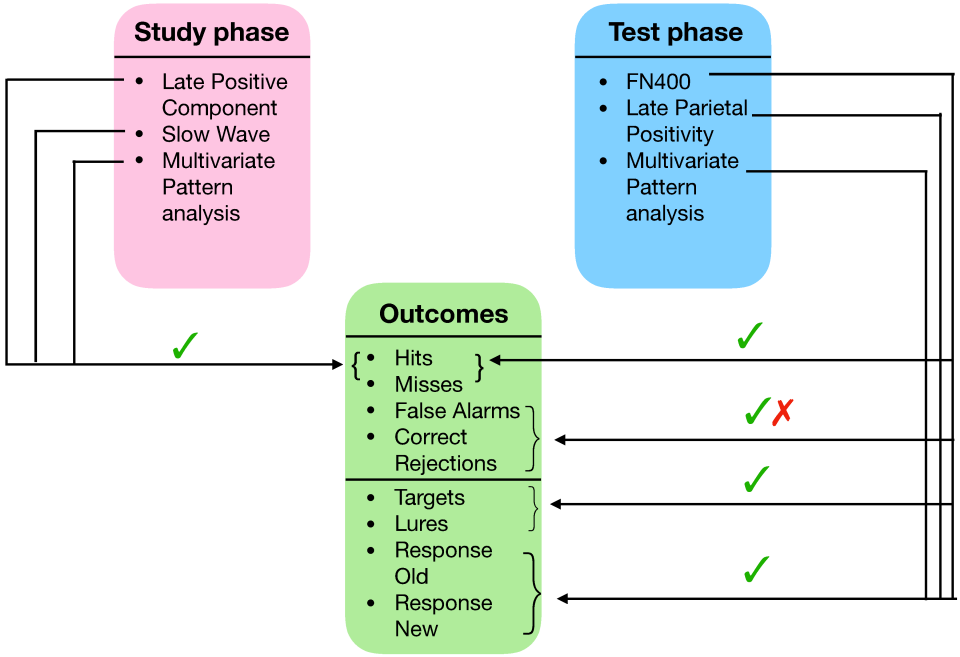

Figure 6.1: Summary of the main investigations pursued in Chapters 2 and 3 based on the item recognition task. Arrows indicate the classification problems of interest and the chosen brain-activity measures used to test for the predictions. Green ticks indicate significant predictive success (across participants). For classification of correct rejections and false alarms, only multivariate pattern analysis of the test-phase brain activity achieved significant success.

(LPC) and slow wave (SW), achieved small but significant success in predicting subsequent memory. This provides objective evidence for the predictability of the subsequent memory effect ERPs (see Figure 2.5 in Chapter 2, page 48). However, across participants, the small size of prediction (estimated from the 95% confidence intervals) could suggest that explaining memory-variability by cognitive processes underlying the LPC and SW amplitudes is an overestimation. Interestingly, predictions based on LPC and SW amplitudes were positively correlated with each other (see Figure 2.6 in Chapter 2, page 48), which contrasts with the general suggestion that LPC and SW index different cognitive processes such as shallow- and deep encoding strategies, respectively (Paller et al., 1987; Smith, 1993). Chen et al. (2014) also found correlations between the trial-averaged amplitudes for LPC and SW (when comparing between subsequent hits and misses). However, only LPC amplitude correlated with participants' d', suggesting a greater relevance of LPC than SW for the memory outcomes of this task. Also, consider that LPC and SW have 1) similar scalp-distribution of voltage, 2) they are generally computed from the same electrode Pz, and 3) their latencies may overlap. Thus, correlations between measures of LPC and SW (average amplitudes or AUCs) could be due to the general temporal and/or spatial auto-correlation property of the EEG recordings. We also found that predictions for hits and misses based on the LPC amplitude correlated with that based on the FN400 amplitude (at test), whereas predictions based on the SW amplitude correlated with that based on the late parietal positivity (LPP) amplitude (at test) (see Figure 3.20 in Chapter 3, page 112). This could suggest a difference between the two ERPs at study. In parallel, Chen et al. (2014) reported correlations between the trial-averaged amplitudes of LPC and FN400, and also between SW and LPP. Together these results suggest that despite some commonality, either caused by the way these two signals are measured or due to their shared variance in explaining subsequent memory success, LPC and SW are, at least in part, signals that index different cognitive processes at study.

FN400, late parietal positivity, and single- versus dual-process accounts The amplitudes of the two commonly-investigated ERPs during the test phase of the item recognition task, FN400 and LPP, also achieved modest success in classifying between 1) targets and lures, 2) hits and misses, and 3) old and new responses. These results support the relevance of the FN400 and LPP in recognition memory outcomes (see Figure 3.2 in Chapter 3, page 87). However, both ERP signals failed to classify between false alarms and correct

rejections, which contrasts with the suggestion that false alarms are processed similar to targets than lures, and are driven by familiarity (Finnigan et al., 2002; Wolk et al., 2006). Unlike LPC and SW at study, predictions based on FN400 and LPP amplitudes did not correlate with each other. Thus, FN400 and LPP may not contribute to common memory variability (see Figure 3.3 in Chapter 3, page 89). There is little doubt that FN400 and LPP are characteristically different, the scalp-distributions of voltage for the two ERPs look clearly very different (see Figure 3.1 in Chapter 3, page 86). Also, the common view is that FN400 and LPP index familiarity- and recollection-based cognitive processes, respectively (Rugg & Curran, 2007). However, debates surround the idea whether FN400 and LPP (or the familiarity and recollection processes that may underlie these signals) *independently* give rise to the recognition judgments, or if they are integrated to form a unitary source of evidence that drives the recognition judgments. The current investigations did not settle this debate, but offered a few useful insights. Specifically, the consideration of response times could be an important aspect. Analysis of the response times for the memory judgments showed that on average, many responses were reached either before or within the time-window of interest for the LPP (see Figure 3.6 in Chapter 3, page 92). Thus, LPP amplitude-based predictions, at least for the shorter response-time trials, may have been supported either by some motor-preparatory activity due to the response-actions, or driven by a recollection process that was more epiphenomenal in this task. Also, Chen et al. (2014) reported that d' correlated with trial-averaged FN400- but not LPP amplitudes. Accordingly, one possible interpretation is that in this task, LPP supported meta-judgments about recently made decisions, rather than actively driving the memory judgments (Woroch & Gonsalves, 2010).

Feedback-related negativity, reward-prediction error, and trial-and-error learning The possible connection with reinforcement learning theories have led to an abundance of studies pursuing the characteristics of the feedback-related negativity (FRN; for a review, see Walsh & Anderson, 2012). Indeed it is an attractive idea that there may exist a signal in the brain that does exactly what a reward-prediction error (RPE) function would do in a theoretical problem (Holroyd & Coles, 2002). However, this may be an oversimplification of the functions of the FRN, specifically if we consider the fact that trial-and-error learning is more basic to the range of cognitive functions that higher-order cortical areas are capable of supporting, and that the FRN is very likely generated in one of the higher-order cortical

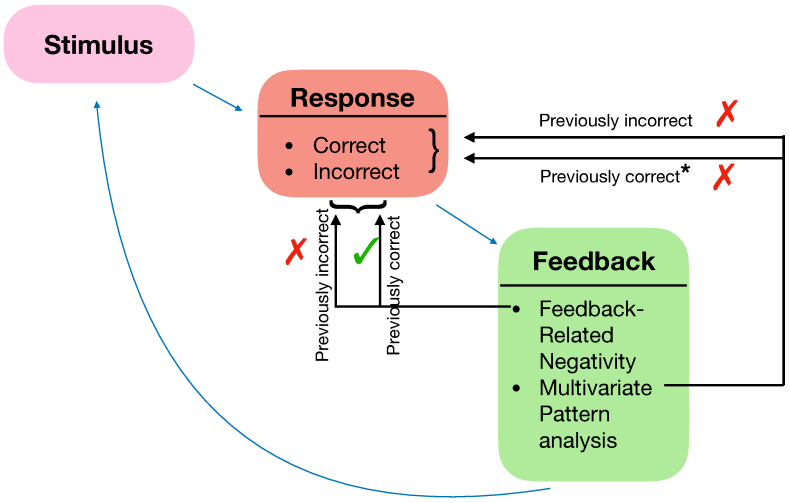

Figure 6.2: Summary of the main investigations pursued in Chapter 5 with the trial-and-error learning task. Blue arrows indicate the flow of the task; black arrows indicate the classification problems of interest and the brain-activity measures used to test for predictions. Green ticks indicate significant predictive success (across participants). Red crosses indicate failure to find a significant effect. *Analysis was substantially under-powered.

areas— the anterior cingulate cortex (ACC). Moreover, recent evidence suggests that for reinforcement learning, the role of the higher-order cortical areas may not be simply limited to relaying of information from the midbrain dopaminergic systems, but they could also play a more active role, specifically when the task involves complex learning variables (Barraclough et al., 2004; Gershman & Daw, 2017). Thus, the finding that the current trial-and-error learning task recruited the FRN differently from what would be expected based on a purely RPE function (Chapter 4), may not be that surprising. However, EEG experiments cannot claim involvement of specific higher-order cortical regions, and thus, future investigations with fMRI, will be required to examine the above suggestion. Also, it is possible that the signal identified with the planned comparisons approach in Chapter 4 is different from the FRN signal suggested by previous studies, especially because its scalp distribution of voltage appeared to be relatively more frontal than the FRN-specific fronto-central negativity.

Given that the FRN is related to how and what information is retrieved from the feedback, it likely plays a role in supporting trial-and-error learning. However, the question of interest for the current study (Chapter 5) was whether or not the FRN supports learning following an RPE-like function. The results did not suggest so. Both the trial-averaged ERP effects and the predictions with the amplitudes of the FRN-like signal indicated that this signal supported how learned (or correct) responses were maintained across successive cycles, rather than how errors were adjusted subsequently (see Figures 5.1 and 5.2 in Chapter 5, pages 177 and 178, respectively). A similar pattern of results was also reported by Arbel et al. (2014), who used a paradigm similar to the current task. The suggestion that FRN supports learning following an RPE-like function, has also been challenged with the more commonly-used gambling paradigm (Chase et al., 2011). Critically departing from the studies above, the results presented in Chapter 5 were from direct tests of predictions based on the amplitude of an FRN-like signal. Notably, visual inspection of the feedback-locked ERPs also showed other characteristics, for example, a sustained SW-like signal (see Figure 5.1 in Chapter 5, page 177), which could have separated the different learning-relevant conditions. Future investigations will be required to carry out a more extensive analysis of the ERP signals present during feedback-processing, and their role in supporting trial-to-trial learning of many stimulus-response rules.

6.1.2 Additional insights from multivariate pattern analysis

Correlation between classifier performance and participants' performance Across Chapters 2, 3 and 5, a common finding was that classifier performance was positively correlated with the participants' performance. For example, Chapter 2 showed that SVM performance for predicting subsequent memory success (hits or misses) was positively correlated with participants' d'. Although the overall size of prediction for SVM was small when predicting subsequent memory, it was meaningfully large for better-performing participants. Likewise, Chapter 3 showed that participants' d' was positively correlated with both LDA and SVM performance for prediction of old and new trials (see Figure 3.10 in Chapter 3, page 97). In Chapter 5, SVM performance for classifying between subsequent correct and incorrect responses, when restricted to previously-incorrect trials, was correlated with the average accuracy of the participant for the last four training cycles Figure 5.5 in Chapter 5, page 181). Taken together, these results could suggest that the probability of success of a classifier may not only depend on the underlying cognitive processes, but also how those are reflected in the recorded brain activity. Participants with better recognition memory may have a higher SNR that could be picked up by the classifiers. Better-performing participants may also be more engaged in the task, which produced brain activity that was more relevant to the task and could be leveraged by the classifiers to make better predictions.

Cognitive processes at study Multivariate pattern analysis of study-phase activity for the item recognition task produced a small significant effect (see Figure 2.8 in Chapter 2, page 49), suggesting that cognitive processes present during the study phase contributed to only a small amount of variability in later memory success. Additionally, with cluster analysis of the LDA weights we found that there were two main classifier-identified patterns, for two different subgroups of participants (see Figure 2.11 in Chapter 2, page 56). Thus, there were possible differences due to how participants approached the study phase. However, as I explain later, since classifier analysis is a data-driven technique, classifier-identified patterns are not necessarily the same as those used by the brain in making the memory judgments.

Single- and dual-process accounts Partly supporting the idea that FN400 and LPP may not contribute to common variability in memory, multivariate pattern analysis of test-phase activity, which could find combinations of features, such as the FN400 and LPP,

produced significantly better predictions than predictions based on FN400 or LPP amplitudes alone (see Figure 3.9 in Chapter 3, page 96). Importantly, unlike FN400 or LPP, multivariate pattern analysis succeeded in predicting false alarms and correct rejections, and the classifier-identified pattern indicated that there may exist other memory-relevant signals beyond FN400 and LPP. Further investigations of the characteristics of memory-relevant signals beyond FN400 and LPP, are important goals for future research. Also, classifier evidence for smaller time-intervals, after correcting for shorter response times with vincentization, added further insights into the single- and dual-process accounts (see Figure 3.17 in Chapter 3, page 108). Specifically, for shorter response-time trials, there was some evidence for a unitary, integrated signal driving the memory judgments, more in line with a single-process account (Dunn, 2008) than a dual-process account (Yonelinas, 2002). However, for longer response-time trials, there was clearly an independent early-signal, and also likely a late signal, contributing to the memory-judgments. Importantly, the signals were not integrated to produce a single evidence, but the influence of the later signal may have been relatively small. Thus, longer response-time trials showed more support for a dual- than a single-process account. Together, these suggest that the single- and dual-process accounts of recognition memory need not be strictly disjoint, and also, can be viewed in terms of the time it takes for the participant to reach a decision. True recollection-based decisions may be rare, at least for making the simple old/new judgments in this task. A recollection-like process is more likely present when it takes longer to reach the decision, but even then, the recollection-dependent evidence may only be partly different from the earlier, familiarity-driven evidence. It is also possible that the LPP is not a good marker of recollection-based recognition.

The remember/know paradigm (Tulving, 1985) has added much support to the interpretations of FN400 and LPP based on familiarity and recollection, respectively, and thereby to a dual-process account. The remember/know paradigm is a derivative of the basic old/new task, where participants are asked if, for the old response, they can 1) recollect specific details about the item or 2) can remember the item without recollecting specific details (in other words, if they can remember the item with a sense of familiarity only), and respond 'remember' or 'know', respectively. Multiple studies have found that in the remember/know paradigm, LPP amplitude is modulated by the difference in remember and know responses (Curran, 1999, 2004; Rugg & Yonelinas, 2003). However, the remember/know paradigm may

specifically prime the participant towards the meta-judgments at a greater level than what happens for the basic old/new judgments, and thus, recruit the recollection process more widely. Moreover, Dunn (2008) argued that even remember/know responses can be based on a single integrated strength. Thus, future investigations using the classifier approach to examine neural evidence for memory relevant information for the remember/know paradigm could be helpful to further understand the role of LPP.

Comparing between cognitive processes at study and at test Based on the prediction size for classifying memory success versus failure (hits and misses) with the univariate ERP amplitudes and the multivariate pattern analysis, for both study and test phase activity, it is possible that cognitive processes present during the test phase are able to retrieve those at study, and also contribute to memory variability on their own. Multiple results supported this idea. First, as mentioned before, the correlations between predictions with LPC (at study) and FN400 (at test) amplitudes, and between SW (at study) and LPP (at test) amplitudes, may suggest commonality in the memory-specific functions of the 1) LPC and FN400 and 2) SW and LPP. Second, multivariate pattern analysis with test-phase activity surpassed all other measures in the predictive strength. This may not be surprising as test-phase activity likely includes more relevant information for the memory judgments, and the classifier results supported this idea. Considering the variability in memory, it may be more surprising that the classifiers were able to find small but significant predictive success with the study-phase activity. Lastly, the study+test classifiers performed similar to the test-phase classifiers (see Figure 3.19 on page 111), suggesting that incorporating both study and test activity does not produce an additive effect. In other words, the classifiers show that the predictive power of study activity is fully absorbed by the test activity; this is an interesting and novel finding.

Classifiers for trial-and-error learning Predicting response-accuracy of a trial from the feedback-locked activity of the previous trial (Chapter 5) was based on the same logic as the subsequent memory effect. Additionally, with the repetitions of the words across the cycles and the use of rewards as reinforcements, the trial-and-error learning task likely engaged the participants better than the item recognition task. However, multivariate pattern analysis of the feedback-locked activity to predict subsequent response-accuracy was limited by the

dependence of accuracy on cycles. Since accuracy increased with cycles, classifiers could use the information about difference due to cycles to predict response accuracy; thus, the classifiers could make predictions without relying on actually task-relevant brain activity.

For predictions with the amplitude of the FRN-like signal, the classification rule was already assumed (based on the RPE function of the FRN), and thus, no separate training trials were required. On the other hand, to learn the classification rules, classifiers (LDA and SVM) required a larger number of training trials; thus, considering a single cycle may not have been helpful. Also, trials from multiple cycles were not pooled together because of the accuracy-cycle dependence. Still, predicting subsequent response-accuracy based on the feedback-locked activity of the previous cycle was significantly above chance, when considering all trials in the previous cycle (see Figures 5.3 and 5.4 in Chapter 5, pages 180 and 180, respectively). However, the classifier could also be predicting the response accuracy of the current cycle, which is correlated to the response accuracy of the next. Considering only-correct or only-incorrect trials in the previous cycle was a cleaner comparison, and was not susceptible to this confound. However, this restriction cut down the number of trials available to the classifier training even more drastically, specifically for the case of previously-correct trials. Thus, in the future, we will need to figure out more clever ways to implement the classifiers for understanding these types of learning situations.

Interestingly, the classifiers achieved modest success in predicting word-value (high or low), which was not confounded by the accuracy-cycle dependence, and was done separately for correct and incorrect feedback outcomes (see Figure 5.6 in Chapter 5, page 183). Further, the classifier-identified patterns were different for when the feedback outcome was correct and when it was incorrect. This could suggest that learning from correct and incorrect feedback outcomes involved different signals or different combinations of the signals present during feedback processing. Notably, the above result also parallels the results from the planned comparisons of the feedback-locked ERPs - the FRN-like signal in cycle 1 indexed subsequent response adjustments in cycle 2 but only for the correct and not the incorrect trials in cycle 1. Thus, another future direction will be to investigate why and how learning from correct and incorrect feedback differed, even though both were equally useful in learning the stimulus-specific response-rules.

6.2 Research significance

To examine the brain-basis of behaviour, there have been two major approaches. One approach is to look into loss of cognitive functions following brain lesions, which provides more causal support for region-specific activity and behaviour. For example, studies with patient H.M., who had undergone surgery to control severe epileptic seizures, provided detailed accounts of brain-regions relevant to memory processes. H.M.'s surgery resulted in lesions to different parts of the medial-temporal lobe, including the hippocampus, the amygdala and the parahippocampal gyrus. Post surgery, and in the following years, H.M. showed severe deficits in remembering new episodic-details with no apparent loss in perceptual or intellectual abilities. This suggested the importance of the medial temporal lobe in memory functions (Scoville & Milner, 1957).

However, conducting studies following brain-lesions is limited to specific patient populations only. The other research theme, that became popular with the development of various techniques to record functional brain-activity, such as, EEG, fMRI etc., has commonly looked into the neural correlates of behaviour. These studies can be conducted with representative samples from the general (without lesion) population. With the superior spatial-resolution, fMRI can be used to identify region-specific activity for psychological processes; whereas with EEG, which has a superior temporal-resolution, researchers have looked into the time-course of brain activity, evoked by psychological processes, and with milliseconds accuracy. However, causality cannot be claimed with this approach. Different psychological processes can recruit the same brain-region in fMRI, or the same EEG signal. Reports of individual differences within the same experiment are also common. Moreover, many researchers suggest that psychological processes are more likely to be produced by distributed pattern of activity in the neuronal-networks than by localized activity.

Predictive approaches cannot claim causality either, but are definitely a step ahead in finding the neural-correlates, than the planned-comparisons and descriptive methods. Specifically, the idea of generalizability is important, for otherwise we may be looking at sample-specific solutions only. Statistical tests for ERP effects are typically conducted for a very small number of electrodes. Evaluating the signals from across the scalp is important, but with the traditional approach, it requires corrections for multiple comparisons, and are not commonly pursued. Multivariate methods, including machine learning classifiers, could be

better suited for evaluating signals across the scalp.

Moreover, classifiers are data-driven techniques, and thus, they do not require a specific hypothesis about the direction of the effect. For example, ERP effects of LPC are evaluated based on the hypothesis that for subsequent hits, LPC amplitude is more positive than that for subsequent misses. However, when classifying subsequent hits and misses based on study-phase activity, the classifiers can automatically learn how brain activity differs for hits and misses, by learning the characteristics of the features specific to subsequent memory success.

The current investigations showed that questions of interest to traditional ERP research can also be addressed with predictive approaches, which add more stringent criteria for evaluating behaviourally-relevant brain activity. The ROC analysis of the ERP amplitudes paralleled the effects with the trial-averaged ERP amplitudes. Although both were significant, the two analyses were characteristically different because the trial-averaged ERP amplitudes cannot indicate the amount of separability between conditions at the individual trial level. Consider that if the means for the two conditions are quite different, the distributions could still largely overlap, leading to poorer predictions for individual trials (see Figure 1.5c on page 20 of the Introduction Chapter). However, the tests for predictions can tell whether or not the trial-averaged ERP amplitudes are different for the pair of conditions. Also, the measures of area under the curve (AUC) of the receiver operating characteristic (ROC) curves are balanced and absolute measures. These can be compared to chance (0.5), as well as compared across different brain activity measures. This is not possible with the trial-averaged ERP effects, which only evaluate relative difference between conditions.

Further, the implementation of predictive approaches for investigating brain activity underlying learning outcomes offered more than a "proof-of-principle" for information in the brain. For example, examination of the dynamics of brain activity prior to reaching the memory decisions for the item recognition task was different from how classifiers have frequently been used to investigate brain activity, such as, to check whether or not predictions are possible. Instead, by evaluating classifier performance as function of time, this investigation offered novel insights about single- and dual-process accounts of recognition memory.

6.3 Limitations and future directions

Despite offering a better lens to evaluate behaviourally-relevant brain activity, some limitations of the predictive approach, specifically in the context of the current investigations, are also apparent.

Availability of a large dataset and sensitivity to circumstantial/artifactual features In general, classifiers require a relatively large set of trials for training, and even more so, when the data include many features; though support vector machine (SVM) models are more robust against overfitting when the number of features is large in comparison to the small number of training trials. Thus, classifiers may not be well-suited for experiments with too few trials. EEG experiments can be designed to include many trials, but longer recording-sessions can also influence the data quality. As I have discussed before, classifiers can be sensitive to such information. For example, in Chapter 2 (Chakravarty et al., 2020), we predicted subsequent memory success from the study-phase activity, and for the individual study lists (see Figure 2.14 in Chapter 2, page 62). This analysis showed that classifier performance increased with the order of the study lists. One possible reason behind this result could be that there were gradual changes in the EEG signal over the recording session, which somehow helped the classifiers to better discriminate between subsequent hits and misses. On the other hand, recordings over multiple sessions can also be problematic, for this could shift the electrode positions, the brain-state, and change the signals thereby.

Further, classifiers can be quite sensitive to circumstantial features. The biggest example for the current investigation was classifying subsequent response accuracy based on the feedback-locked activity from the previous trial (Chapter 5). Due to being data-driven techniques, the classifiers need not rely on task-relevant brain activity to predict the behaviour. Instead, they can pick up on any information that is useful for such prediction. Specifically when the signal-changes for the task-relevant brain activity are smaller than the differences indexed by circumstantial features, it is important to carefully review the question of interest for the classifier investigation, in order to make meaningful inferences from these methods. However, if the signal changes for the the task-relevant brain activity are not that small, circumstantial features may not influence the classifiers as much. One example of this was the motor preparatory activity present during test-phase of the item recognition task, which

did not seem to have a significant impact on the predictions (Chapter 3).

Difficulty interpreting classification results It cannot be determined in advance as to which classifier models should be used; the process of selecting a classifier is exploratory. The current investigations included the two most simple classifiers, linear discriminant analysis (LDA) and SVM. There were two reasons behind this selection. First, the objective was to evaluate the general level of challenge in predicting learning-success from brain-activity measures, with possible methodological improvements kept as a separate goal for the future. Second, non-linear classifiers, which usually evaluate independent contributions of its parameters, along with their interactions, could be better suited to situations where larger sets of trials are available to train the classifiers. However, it is possible that learning outcomes are represented in brain activity patterns that are not linearly separable, and would need to be investigated in the future.

Non-reliance on theory Lastly, even when it is not influenced by circumstantial features, a classifier can find combinations of features (to predict behaviour) that is different from the pattern of activity used by the brain to produce the same behaviour. With LDA, it was possible to take a direct peak at the features with greater influence and compare those with existing findings from ERP research. However, in general, interpreting the features for the classifier analysis, specifically for non-linear classifiers is not straightforward. Thus, the lack of dependence on existing theories is both an advantage and a disadvantage for the classifier methods. Accordingly, a better approach is to incorporate insights from both planned-comparisons approaches and the predictive methods.

In view of the current results, there are several directions for future research using the classifier methods.

Manipulation of engagement in the task As mentioned before, across multiple analyses, classifier performance correlated with participants performance, suggesting that better-performing participants, who may have a higher task-engagement or motivation to do well in the task, could lead to capturing of more task-relevant brain activity by EEG recordings, which in turn could be picked up by the classifiers. Accordingly, a future manipulation of task engagement could lead to overall better predictions. However, depending on the

classification problem of interest, better-performing participants could also produce greater imbalance between the classes (e.g., hits and misses), which would need to be accounted for.

Leveraging the greater signal-to-noise ratio of the participants who perform well Following up on the above idea, another future direction can be investigated. Consider that for the better-performing participants, the signal-to-noise ratio (SNR) may be relatively higher for the classifiers to pick up on. Then, we can ask if it is possible to make predictions about behavioural outcomes for the other participants, based on the training received from the brain-activity for the better-performing participants. The non-stationarity of the signal due to the potentially different states of the different participants, would need to be accounted for, but our initial investigations suggest that this between-subject classification analysis could be successful (e.g., see Figure 2.15 in Chapter 2, page 65).

Learning applications based on brain activity Predicting memory from brain activity that precedes the memory tests has important learning applications. For example, online predictions of memory from the study-phase activity could be used to train participants to study better. Further, with techniques like neurofeedback protocols, these predictions can be used to help participants self-regulate into states that are more conducive of memory success (for an implementation of classifier-driven neuro-stimulation, see Ezzyat et al., 2017). Also, the data-driven approach of this research makes it equally amenable to the normal, healthy population, as well as specific patient populations.

6.4 Conclusion

In sum, this work shows that brain activity, underlying learning outcomes, can be investigated using predictive approaches, which impose a more stringent criteria for evaluating brain-activity as behaviourally-relevant. I showed that questions of interest to traditional ERP research, are also of relevance, and can be tested, with the predictive methods. Additionally, multivariate pattern analysis may identify behaviourally-relevant signals that are not commonly studied with the traditional ERP methods. Moreover, the size of prediction can be used to compare between the contribution of different signals in explaining the learning-variability; this is not possible with the traditional ERP methods. Also, unlike univariate ERP measures, the classifiers can be set up to examine dynamics of brain activity

leading up to memory decisions. For the current investigations, the overall size of prediction with the time-domain EEG features was small, which could be what we can generally expect to find from predictive analysis of brain-activity for different behavioural outcomes. Aside from potential improvements to the classifier analysis discussed above, the small size of prediction could suggest that the resolution of the EEG recordings may not be enough to capture all relevant factors influencing learning outcomes, and interestingly, this resolution could depend on the participants' level of engagement in the task. Future investigations will be required to test these suggestions in detail.

Bibliography

Ally, B. A., Simons, J. S., McKeever, J. D., Peers, P. V., & Budson, A. E. (2008). Parietal contributions to recollection: electrophysiological evidence from aging and patients with parietal lesions. *Neuropsychologia*, *46*(7), 1800–1812.

Arbel, Y., Goforth, K., & Donchin, E. (2013). The good, the bad, or the useful? the examination of the relationship between the feedback-related negativity (FRN) and long-term learning outcomes. *Journal of Cognitive Neuroscience*, *25*(8), 1249–1260.

Arbel, Y., Murphy, A., & Donchin, E. (2014). On the utility of positive and negative feedback in a paired-associate learning task. *Journal of Cognitive Neuroscience*, *26*(7), 1445–1453.

Arbel, Y., & Wu, H. (2016). A neurophysiological examination of quality of learning in a feedback-based learning task. *Neuropsychologia*, *93*, 13–20.

Arora, A., Lin, J.-J., Gasperian, A., Maldjian, J., Stein, J., Kahana, M., & Lega, B. (2018). Comparison of logistic regression, support vector machines, and deep learning classifiers for predicting memory encoding success using human intracranial EEG recordings. *Journal of Neural Engineering*, *15*(6), 066028.

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, *7*(4), 404–410.

Bellebaum, C., & Daum, I. (2008). Learning-related changes in reward expectancy are reflected in the feedback-related negativity. *European Journal of Neuroscience*, *27*(7), 1823–1835.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for english words (ANEW): Instruction manual and affective ratings* (Tech. Rep.). : Technical report C-1, the Center for Research in Psychophysiology.

Brewer, J. B., Zhao, Z., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. (1998). Making memories: brain activity that predicts how well visual experience will be remembered. *Science*, *281*(5380), 1185–1187.

Cabeza, R., Ciaramelli, E., Olson, I. R., & Moscovitch, M. (2008). The parietal cortex and episodic memory: an attentional account. *Nature reviews neuroscience*, *9*(8), 613–625.

Chakravarty, S., Chen, Y. Y., & Caplan, J. B. (2020). Predicting memory from study-related brain activity. *Journal of Neurophysiology*, *124*(6), 2060–2075.

Chakravarty, S., Fujiwara, E., Madan, C. R., Tomlinson, S. E., Ober, I., & Caplan, J. B. (2019). Value bias of verbal memory. *Journal of Memory and Language*, *107*, 25–39.

Chase, H. W., Swainson, R., Durham, L., Benham, L., R., & Cools. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic

reversal learning. *Journal of Cognitive Neuroscience*, *23*, 936–946.

Chavarriaga, R., Sobolewski, A., & Millán, J. d. R. (2014). Errare machinale est: the use of error-related potentials in brain-machine interfaces. *Frontiers in neuroscience*, *8*, 208.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, Y. Y., Lithgow, K., Hemmerich, J. A., & Caplan, J. B. (2014). Is what goes in what comes out? encoding and retrieval event-related potentials together determine memory outcome. *Experimental Brain Research*, *232*(10), 3175–3190.

Cockburn, J., & Holroyd, C. B. (2018). Feedback information and the reward positivity. *International Journal of Psychophysiology*, *132*, 243–251.

Cohen, M. X., Elger, C. E., & Ranganath, C. (2007). Reward expectation modulates feedback-related negativity and EEG spectra. *Neuroimage*, *35*(2), 968–978.

Collins, A. G., & 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*, *115*(10), 2502–2507.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Coutanche, M. N. (2013). Distinguishing multi-voxel patterns and mean activation: why, how, and what does it tell us? *Cognitive, Affective, & Behavioral Neuroscience*, *13*(3), 667–673.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

Criss, A. H., Wheeler, M. E., & McClelland, J. L. (2013). A differentiation account of recognition memory: evidence from fmri. *Journal of Cognitive Neuroscience*, *25*(3), 421–435.

Curran, T. (1999). The electrophysiology of incidental and intentionalretrieval: ERP old/new effects in lexical decision and recognition memory. *Neuropsychologia*, *37*(7), 771–785.

Curran, T. (2004). Effects of attention and confidence on the hypothesized ERP correlates of recollection and familiarity. *Neuropsychologia*, *42*(8), 1088–1106.

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, *319*(5867), 1264–1267.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21.

Diedrich, O., Naumann, E., Maier, S., Becker, G., & Bartussek, D. (1997). A frontal positive slow wave in the erp associated with emotional slides. *Journal of Psychophysiology*, *11*, 71–84.

Dolcos, F., LaBar, K. S., & Cabeza, R. (2005). Remembering one year later: role of the amygdala and the medial temporal lobe memory system in retrieving emotional memories. *Proceedings of the National Academy of Sciences*, *102*(7), 2626–2631.

Dunn, J. C. (2008). The dimensionality of the remember–know task: a state-trace analysis. *Psychological Review*, *115*(2), 426-446.

Eldridge, L. L., Sarfatti, S., & Knowlton, B. J. (2002). The effect of testing procedure on remember–know judgments. *Psychonomic Bulletin & Review*, *9*(1), 139-145.

Ernst, B., & Steinhauser, M. (2012). Feedback-related brain activity predicts learning from feedback in multiple-choice testing. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(2), 323–336.

Ezzyat, Y., Kragel, J. E., Burke, J. F., Levy, D. F., Lyalenko, A., Wanda, P., ... Kahana, M. J. (2017). Direct brain stimulation modulates encoding states and memory performance in humans. *Current Biology*, *27*(9), 1251–1258.

Fabiani, M., Karis, D., & Donchin, E. (1990). Effects of mnemonic strategy manipulation in a Von Restorff paradigm. *Electroencephalography and Clinical Neurophysiology*, *75*(1-2), 22–35.

Finnigan, S., Humphreys, M. S., Dennis, S., & Geffen, G. (2002). Erp 'old/new' effects: memory strength and decisional factor (s). *Neuropsychologia*, *40*(13), 2288–2304.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Frank, M. J., Woroch, B. S., & Curran, T. (2005). Error-related negativity predicts reinforcement learning and conflict biases. *Neuron*, *47*, 495–501.

Friedman, D. (1990). Cognitive event-related potential components during continuous recognition memory for pictures. *Psychophysiology*, *27*(2), 136–148.

Fukuda, K., & Woodman, G. F. (2015). Predicting and improving recognition memory using multiple electrophysiological signals in real time. *Psychological Science*, *26*(7), 1026–1037.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system of error detection and compensation. *Psychological Science*, *4*, 385–390.

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, *295*, 2279–2282.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual Review of Psychology*, *68*, 101–128.

Goyer, J. P., Woldorff, M. G., & Huettel, S. A. (2008). Rapid electrophysiological brain responses are influenced by both valence and magnitude of monetary rewards. *Journal of Cognitive Neuroscience*, *20*, 2058–2069.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics* (Vol. 1). Wiley New York.

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language*, *94*, 149–165.

Guo, C., Duan, L., Li, W., & Paller, K. A. (2006). Distinguishing source memory and item memory: Brain potentials at encoding and retrieval. *Brain Research*, *1118*(1), 142–154.

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, *71*(2), 148–154.

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2007). It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology*, *44*(6), 905–912.

Halpern, D., Tubridy, S., Wang, H. Y., Gasser, C., Popp, P. O., Davachi, L., & Gureckis, T. M. (2018). Knowledge tracing using the brain. *International Educational Data Mining Society*.

Hart, A. S., Rutledge, R. B., Glimcher, P. W., & Phillips, P. E. (2014). Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *Journal of Neuroscience*, *34*(3), 698–704.

Hauser, T. U., Iannaccone, R., Stämpfli, P., Drechsler, R., Brandeis, D., Walitza, S., & Brem, S. (2014). The feedback-related negativity (FRN) revisited: new insights into the localization, meaning and network organization. *Neuroimage*, *84*, 159–168.

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

Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? *The Quarterly Journal of Experimental Psychology Section A*, *58*(2), 193–233.

Hicks, J. L., & Marsh, R. L. (1999). Remember–know judgments can depend on how memory is tested. *Psychonomic Bulletin & Review*, *6*(1), 117-122.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*, 679–709.

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R. B., & Coles, M. G. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature Neuroscience*, *7*, 497–498.

Holroyd, C. B., Pakzad-Vaezi, K. L., & Krigolson, O. E. (2008). The feedback correct-related positivity: sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology*, *45*, 688–697.

Höltje, G., & Mecklinger, A. (2018). Electrophysiological reward signals predict episodic memory for immediate and delayed positive feedback events. *Brain Research*, *1701*, 64–74.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: a theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*(2), 208.

Jamalabadi, H., Alizadeh, S., Schönauer, M., Leibold, C., & Gais, S. (2016). Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. *Human Brain Mapping*, *37*(5), 1842–1855.

Kahana, M. J. (2012). *Foundations of Human Memory*. Oxford University Press, USA.

Kamp, S.-M., Bader, R., & Mecklinger, A. (2017). ERP subsequent memory effects differ between inter-item and unitization encoding tasks. *Frontiers in Human Neuroscience*, *11*, 30.

Karis, D., Bashore, T., Fabiani, M., & Donchin, E. (1982). P300 and memory. *Psychophysiology*, *19*(3), 328–328.

Karis, D., Fabiani, M., & Donchin, E. (1984). "P300" and memory: Individual differences in the Von Restorff effect. *Cognitive Psychology*, *16*(2), 177–216.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, *303*(5660), 1023–1026.

Kim, A. S., Vallesi, A., Picton, T. W., & Tulving, E. (2009). Cognitive association formation in episodic memory: Evidence from event-related potentials. *Neuropsychologia*, *47*(14), 3162–3173.

Kim, H. (2011). Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies. *NeuroImage*, *54*(3), 2446–2461.

Koch, G. E., Paulus, J. P., & Coutanche, M. N. (2020). Neural patterns are more similar across individuals during successful memory encoding than during failed memory encoding. *Cerebral Cortex*, *30*(7), 3872–3883.

LaRocque, K. F., Smith, M. E., Carr, V. A., Witthoft, N., Grill-Spector, K., & Wagner, A. D. (2013). Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. *Journal of Neuroscience*, *33*(13), 5466–5474.

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, *84*(3), 555–579.

Lee, A. C., Brodersen, K. H., & Rudebeck, S. R. (2013). Disentangling spatial perception and spatial memory in the hippocampus: a univariate and multivariate pattern analysis fmri study. *Journal of Cognitive Neuroscience*, *25*(4), 534–546.

Lewis, D. J. (1979). Psychobiology of active and inactive memory. *Psychological Bulletin*, *86*(5), 1054.

Liao, K., Mollison, M. V., Curran, T., & de Sa, V. R. (2018). Single-trial EEG predicts memory retrieval using leave-one-subject-out classification. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 2613–2620).

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.

Luft, C. D. B. (2014). Learning from feedback: the neural mechanisms of feedback processing facilitating better performance. *Behavioural Brain Research*, *261*, 356–368.

Luft, C. D. B., Takase, E., & Bhattacharya, J. (2014). Processing graded feedback: electrophysiological correlates of learning from small and large errors. *Journal of cognitive neuroscience*, *26*(5), 1180–1193.

Madan, C. R., Fujiwara, E., Gerson, B. C., & Caplan, J. B. (2012). High reward makes items easier to remember, but harder to bind to a new temporal context. *Frontiers in Integrative Neuroscience*, *6*, 61.

Marco-Pallares, J., Cucurell, D., Münte, T. F., Strien, N., & Rodriguez-Fornells, A. (2011). On the number of trials needed for a stable feedback-related negativity. *Psychophysiology*, *48*(6), 852–860.

Martin, L. E., Potts, G. F., Burton, P. C., & Montague, P. R. (2009). Electrophysiological and hemodynamic responses to reward prediction violation. *Neuroreport*, *20*(13), 1140.

Martinez, W. L., Martinez, A. R., & Solka, J. (2017). *Exploratory Data Analysis with MATLAB*. Chapman and Hall/CRC.

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a "generic" neural system for error detection. *Journal of Cognitive Neuroscience*, *9*, 788–798.

Moran, T. P., Jendrusina, A. A., & Moser, J. S. (2013). The psychometric properties of the late positive potential during emotion processing and regulation. *Brain research*, *1516*, 66–75.

Mousavi, M., & de Sa, V. R. (2019). Spatio-temporal analysis of error-related brain activity in active and passive brain–computer interfaces. *Brain-Computer Interfaces*, *6*(4), 118–127.

Nairne, J. S. (2002). The myth of the encoding-retrieval match. *Memory*, *10*(5-6), 389–395.

Neath, I. (1998). *Human memory: An introduction to research, data, and theory.* Thomson Brooks/Cole Publishing Co.

Neville, H. J., Kutas, M., Chesney, G., & Schmidt, A. L. (1986). Event-related brain potentials during initial encoding and recognition memory of congruous and incongruous words. *Journal of Memory and Language*, *25*(1), 75–92.

Nieuwenhuis, S., Holroyd, C. B., Mol, N., & Coles, M. G. (2004). Reinforcement-related brain potentials from medial frontal cortex: origins and functional significance. *Neuroscience & Biobehavioral Reviews*, *28*(4), 441–448.

Nieuwenhuis, S., Ridderinkhof, K. R., Talsma, D., Coles, M. G., Holroyd, C. B., & Kok, A. (2002). A computational account of altered error processing in older age: dopamine and the error-related negativity. *Cognitive Affective & Behavioural Neuroscience*, *2*, 19–36.

Nieuwenhuis, S., Slagter, H. A., Von Geusau, N. J. A., Heslenfeld, D. J., & Holroyd, C. B. (2005). Knowing good from bad: differential activation of human cortical areas by positive and negative outcomes. *European Journal of Neuroscience*, *21*(11), 3161–3168.

Noh, E., Herzmann, G., Curran, T., & de Sa, V. R. (2014). Using single-trial EEG to predict and analyze subsequent memory. *NeuroImage*, *84*, 712–723.

Noh, E., Liao, K., Mollison, M. V., Curran, T., & de Sa, V. R. (2018). Single-trial EEG analysis predicts memory retrieval and reveals source-dependent differences. *Frontiers in Human Neuroscience*, *12*, 258.

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

Oliveira, F. T., McDonald, J. J., & Goodman, D. (2007). Performance monitoring in the anterior cingulate is not all error related: expectancy deviation and the representation of action-outcome associations. *Journal of Cognitive Neuroscience*, *19*, 1994–2004.

Otten, L. J., & Rugg, M. D. (2001). Electrophysiological correlates of memory encoding are task-dependent. *Cognitive Brain Research*, *12*(1), 11–18.

Paller, K. A. (1990). Recall and stem-completion priming have different electrophysiological correlates and are modified differentially by directed forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(6), 1021.

Paller, K. A., Kutas, M., & Mayes, A. R. (1987). Neural correlates of encoding in an incidental learning paradigm. *Electroencephalography & Clinical Neurophysiology*, *67*(4), 360–371.

Paller, K. A., & Wagner, A. D. (2002). Observing the transformation of experience into memory. *Trends In Cognitive Sciences*, *6*(2), 93–102.

Park, H., & Rugg, M. D. (2010). Prestimulus hippocampal activity predicts later recollection. *Hippocampus*, *20*(1), 24–28.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045.

Pfabigan, D. M., Seidel, E.-M., Paul, K., Grahl, A., Sailer, U., Lanzenberger, R., ... Lamm, C. (2015). Context-sensitivity of the feedback-related negativity for zero-value feedback outcomes. *Biological Psychology*, *104*, 184–192.

Polich, J. (2007). Updating p300: an integrative theory of p3a and p3b. *Clinical neurophysiology*, *118*(10), 2128–2148.

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*(5756), 1963–1966.

Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, *52*(4), 449–459.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*(3), 446.

Ratcliff, R., Sederberg, P. B., Smith, T. A., & Childers, R. (2016). A single trial analysis of EEG in recognition memory: Tracking the neural correlates of memory strength. *Neuropsychologia*, *93*, 128–141.

Rmus, M., McDougle, S. D., & Collins, A. G. (2021). The role of executive function in shaping reinforcement learning. *Current Opinion in Behavioral Sciences*, *38*, 66–73.

Rolls, E. T., McCabe, C., & Redoute, J. (2008). Expected value, reward outcome, and temporal difference error representations in a probabilistic decision task. *Cerebral Cortex*, *18*(3), 652–663.

Rugg, M. D. (1995). *ERP Studies of Memory.* Oxford University Press.

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences*, *11*(6), 251–257.

Rugg, M. D., & Nagy, M. E. (1989). Event-related potentials and recognition memory for words. *Electroencephalography and Clinical Neurophysiology*, *72*(5), 395–406.

Rugg, M. D., & Yonelinas, A. P. (2003). Human recognition memory: a cognitive neuroscience perspective. *Trends in Cognitive Sciences*, *7*(7), 313–319.

Rushby, J. A., Barry, R. J., & Johnstone, S. J. (2002). Event-related potential correlates of serial-position effects during an elaborative memory test. *International Journal of Psychophysiology*, *46*, 13-27.

Rushworth, M. F., & Behrens, T. E. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, *11*(4), 389–397.

Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin*, *141*(1), 213.

SamPenDu, D. (2015). Bayes factors MATLAB functions. .

Sanquist, T. F., Rohrbaugh, J. W., Syndulko, K., & Lindsley, D. B. (1980). Electrocortical signs of levels of processing: perceptual analysis and recognition memory. *Psychophysiology*, *17*(6), 568–576.

Schultz, W. (2015). Neuronal reward and decision signals: from theories to data. *Physiological Reviews*, *95*(3), 853–951.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*(1), 11.

Seamans, J. K., & Yang, C. R. (2004). The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in Neurobiology*, *74*(1), 1–58.

Sederberg, P. B., Gauthier, L. V., Terushkin, V., Miller, J. F., Barnathan, J. A., & Kahana, M. J. (2006). Oscillatory correlates of the primacy effect in episodic memory. *NeuroImage*, *32*, 1422-1431.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, *25*(3), 289–310.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of experimental psychology: General*, *143*(2), 534.

Simon-Thomas, E. R., Role, K. O., & Knight, R. T. (2005). Behavioral and electrophysiological evidence of a right hemisphere bias for the influence of negative emotion on higher cognition. *Journal of Cognitive Neuroscience*, *17*(3), 518–529.

Skocik, M., Collins, J., Callahan-Flintoft, C., Bowman, H., & Wyble, B. (2016). I tried a bunch of things: the dangers of unexpected overfitting in classification. *BioRxiv*, 078816.

Smith, M. E. (1993). Neurophysiological manifestations of recollective experience during recognition memory judgments. *Journal of Cognitive Neuroscience*, *5*(1), 1–13.

Summerfield, C., & Mangels, J. A. (2006). Dissociable neural mechanisms for encoding predictable and unpredictable events. *Journal of Cognitive Neuroscience*, *18*(7), 1120–1132.

Sun, X., Qian, C., Chen, Z., Wu, Z., Luo, B., & Pan, G. (2016). Remembered or forgotten?—An EEG-based computational prediction approach. *PloS one*, *11*(12), e0167497.

Surprenant, A. M., & Neath, I. (2013). *Principles of Memory*. Psychology Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, Mass.

Talmi, D., Grady, C. L., Goshen-Gottstein, Y., & Moscovitch, M. (2005). Neuroimaging the serial position curve: a test of single-store versus dual-store models. *Psychological Science*, *16*(9), 716–723.

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, *307*(5715), 1642–1645.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie canadienne*, *26*(1), 1.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352.

van der Helden, J., Boksem, M. A., & Blom, J. H. (2010). The importance of failure: feedback-related negativity predicts motor learning efficiency. *Cerebral Cortex*, *20*(7), 1596–1603.

van de Vijver, I., Ridderinkhof, K., & Cohen, M. X. (2011). Frontal oscillatory dynamics predict feedback learning and action adjustment. *Journal of Cognitive Neuroscience*, *23*, 4106–4121.

Van Petten, C., & Senkfor, A. J. (1996). Memory for words and novel visual patterns: Repetition, recognition, and encoding effects in the event-related brain potential. *Psychophysiology*, *33*(5), 491–506.

Van Veen, V., Holroyd, C. B., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2004). Errors without conflict: implications for performance monitoring theories of anterior cingulate cortex. *Brain and Cognition*, *56*(2), 267–276.

van Vugt, M., Brandt, A., & Schulze-Bonhage, A. (2017). Tracking perceptual and memory decisions by decoding brain activity. In *Benelux Conference on Artificial Intelligence* (pp. 76–85).

Voss, J. L., Lucas, H. D., & Paller, K. A. (2012). More than a feeling: pervasive influences of memory without awareness of retrieval. *Cognitive Neuroscience*, *3*(3-4), 193–207.

Voss, J. L., & Paller, K. A. (2008). Brain substrates of implicit and explicit memory: The importance of concurrently acquired neural signals of both memory types. *Neuropsychologia*, *46*(13), 3021–3029.

Voss, J. L., & Paller, K. A. (2009). Remembering and knowing: electrophysiological distinctions at encoding but not retrieval. *NeuroImage*, *46*, 280-289.

Wagner, A. D., Koutstaal, W., & Schacter, D. L. (1999). When encodong yields remembering: insights from event-related neuroimaging. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *354*(1387), 1307–1324.

Wagner, A. D., Schacter, D. L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., ... Buckner, R. L. (1998). Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. *Science*, *281*(5380), 1188–1191.

Wagner, A. D., Shannon, B. J., Kahn, I., & Buckner, R. L. (2005). Parietal lobe contributions to episodic memory retrieval. *Trends in Cognitive Sciences*, *9*(9), 445–453.

Walsh, M. M., & Anderson, J. R. (2011a). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences U.S.A.*, *108*, 19048–19053.

Walsh, M. M., & Anderson, J. R. (2011b). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences*, *108*(47), 19048–19053.

Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: Event related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience and Biobehavioral Reviews*, *36*, 1870–1884.

Warren, L. R. (1980). Evoked potential correlates of recognition memory. *Biological Psychology*, *11*(1), 21–35.

Watanabe, T., Hirose, S., Wada, H., Katsura, M., Chikazoe, J., Jimura, K., ... Konishi, S. (2011). Prediction of subsequent recognition performance using brain activity in the medial temporal lobe. *NeuroImage*, *54*(4), 3085–3092.

Weidemann, C. T., & Kahana, M. J. (2019a). Dynamics of brain activity reveal a unitary recognition signal. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(3), 440.

Weidemann, C. T., & Kahana, M. J. (2019b). Neural measures of subsequent memory reflect endogenous variability in cognitive function. *BioRxiv*, 576173.

Weidemann, C. T., Kragel, J. E., Lega, B. C., Worrell, G. A., Sperling, M. R., Sharan, A. D., ... Kahana, M. J. (2019). Neural activity reveals interactions between episodic and semantic memory systems during retrieval. *Journal of Experimental Psychology: General*, *148*(1), 1.

Wilding, E. L., & Rugg, M. D. (1996). An event-related potential study of recognition memory with and without retrieval of source. *Brain*, *119*(3), 889–905.

Williams, C. C., Hassall, C. D., Lindenbach, T., & Krigolson, O. E. (2019). Reward prediction errors reflect an underlying learning process that parallels behavioural adaptations: a trial-to-trial analysis. *Computational Brain & Behavior*, 1–11.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods*, *20*(1), 6–10.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review*, *117*(4), 1025.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, *11*(4), 616–641.

Wolk, D. A., Schacter, D. L., Lygizos, M., Sen, N. M., Holcomb, P. J., Daffner, K. R., & Budson, A. E. (2006). ERP correlates of recognition memory: Effects of retention interval and false alarms. *Brain Research*, *1096*(1), 148–162.

Woroch, B., & Gonsalves, B. D. (2010). Event-related potential correlates of item and source memory strength. *Brain Research*, *1317*, 180–191.

Wu, Y., & Zhou, X. (2009). The P300 and reward valence, magnitude, and expectancy in outcome evaluation. *Brain Research*, *1286*, 114–122.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Yasuda, A., Sato, A., Miyawaki, K., Kumano, H., & Kuboki, T. (2004). Error-related negativity reflects detection of negative reward prediction error. *Neuroreport*, *15*(16), 2561–2565.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological Review*, *111*, 931–959.

Yeung, N., Holroyd, C. B., & Cohen, J. D. (2004). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cerebral Cortex*, *15*(5), 535–544.

Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience*, *24*(28), 6258–6264.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, *25*(6), 747–763.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517.

Yu, R., Zhou, W., & Zhou, X. (2011). Rapid processing of both reward probability and reward uncertainty in the human anterior cingulate cortex. *PloS ONE*, *6*.