Classification of Cognitive Strategies by the underlying processing stages using Hidden semi-Markov Models

Ernö Groeneweg, 4279662
Supervisor/first examiner: Dr. L. van Maanen
Second examiner: Prof. Dr. S. van der Stigchel

MSc Artificial Intelligence

Graduate School of Natural Sciences
Utrecht University
March 2021

Abstract

When doing a cognitive task, people can employ different cognitive strategies. A strategy consists of different cognitive processing stages, which is how different strategies can be differentiated. A novel machine learning method developed by Anderson et al., 2016 is able to model cognitive processing stages in EEG, MEG, and fMRI as a hidden semi-Markov model, calling it Hidden semi-Markov Model Multivariate Pattern Analysis (HsMM-MVPA). This method works across subjects, so among other things it seems to be able to deal with the inter-subject variability of EEG data. This leads to the hypothesis that HsMM-MVPA could potentially be used to predict what cognitive strategy someone used in new, unseen data. To test this hypothesis, EEG data collected from a group of subjects who performed a multiplication task with self-reported cognitive strategies was used. Subjects reported either knowing the answer to a multiplication problem from memory ("retrieval"), or had to compute the answer ("procedural"). We estimated hidden semi-Markov models on some of the subjects and tested how well these models could predict what strategy was used on the other subjects. The models are able to correctly identify retrieval-strategies, but tend to be less sensitive to the procedural-class. This seems to be because the retrieval-strategy is more consistent. HsMM-MVPA can be used for classification, but might fare better with more consistent cognitive strategies.

1 Introduction

Cognitive processing stages make up a central concept in cognitive science (Donders, 1868; Sternberg, 1969). Cognitive processing is a general term used to describe a series of cognitive operations carried out in the creation and manipulation of mental representations of information (Pinheiro-Chagas et al., 2019). Typically, processing stages are applied to behavioural responses. For instance, the classic working memory task by Sternberg, 1969, involves presenting a list of items to memorise, followed by a memory maintenance period during which the subject must maintain the list of items in memory. The maintenance period is terminated by the onset of a 'probe' letter, to which the subject must respond whether the item was in their memorized list of items or not (Sternberg, 1969). Discovering processing stages in the Sternberg task is not complicated, but it has proven a challenge to get exact insights on the duration and temporal onset of individual stages (Henson, 2011, Posner, 2005). In other words, it is still unclear what people are doing exactly when engaging in cognitive processing. Different influential proposals such as sequential processing stage modelling (Zylberberg et al., 2011) or systems factorial technology (Harding et al., 2016) have been proposed, but it seems that response times are not enough to disentangle the onset/offset of processing stages. A recently proposed method of analysing cognitive processing stages by Anderson et al., 2016, called Hidden semi-Markov Model multivariate pattern analysis (HsMM-MVPA), has been suggested to overcome these limits of behavioural data analysis and give more insight into the specifics of processing stages. In this method, cognitive processing stages are modelled as a hidden Markov chain using distributed peaks in activity across an electroencephalographic (EEG) or magnetoencephalographic (MEG) signal. This method has also been successful with functional magnetic resonance imaging (fMRI) analysis (Anderson et al., 2014). By modelling the cognitive processing stages as a semi-Markov chain, we can get insight into the temporal on/offset and durations of processing stages. In this project, we used HsMM-MVPA to build a model-based classifier, to predict which cognitive strategy is used when solving mathematical problems by way of the processing stages that underlie that strategy.

1.1 HsMM-MVPA

HsMM-MVPA was proposed by Anderson et al., 2016. As a method, it is well-suited to parsing a cognitive task into processing stages based on the EEG/MEG samples (Anderson et al., 2016) or fMRI imaging (Anderson et al., 2014) within a trial, instead of using behavioural measures. For example, Anderson et al. found that HsMM-MVPA models could be fitted to data from the original Sternberg task to explain five different stages: (1) preattention, (2) encoding, (3) memory retrieval, (4) decision, and (5) response (Anderson et al., 2016). The method has been shown to be a versatile method for detecting processing stages in a variety of conditions and tasks (Imani et al., 2020; Anderson et al., 2018; Portoles et al., 2018; Zhang, van Vugt, et al., 2018; Zhang, Walsh, et al., 2017; Zhang, Walsh, et al., 2018; Walsh et al., 2017; Zhang, Borst, et al., 2017; Borst and Anderson, 2015; Berberyan et al., 2021). In many implementations, the method can fit models that explain EEG data from a variety of subjects very well. This suggests that there could be some commonality between subjects in how the HsMM-MVPA method represents these processing stages.

Since HsMM-MVPA is able to fit models to data from many subjects at once, it stands to reason that there is enough overlap in cognitive processing stages between subjects employing the same cognitive strategy for the same task. In this project, we set out to discover whether an HsMM-MVPA model can be used to distinguish between cognitive strategies in EEG data from unseen subjects. This means that we have some EEG data of people who are using one of a set of cognitive strategies, but we do not know which one they used and when. We hypothesise that a classifier based on HsMM-MVPA could predict which strategy these unseen subjects used. Using EEG data from subjects doing multiplication sums and then reporting which of two cognitive strategies they used, we used these self-reported strategies as a ground truth to fit Hidden semi-Markov Models to. The best fitting models were then used to estimate the likelihood that a set of unseen trials from entirely new subjects were of one of the strategies or the other. If this labelling performs better than random, then we can say that there is some commonality between processing stages of different individuals employing the same cognitive strategy, which in turn means that their brains are doing something similar.

In a classic Hidden Markov Model (HMM) there are two stochastic finite-time chains of events, one is a hidden Markov chain X and one is an observable chain Y whose behaviour depends on X in some way. For every pair (x, y) where $x \in X$, $y \in Y$ there is an emission probability that x happens when y is observed (Visser et al., 2009). In a classic HMM, the duration of every state corresponds to the duration of a single observation. In a hidden semi-Markov model, we can have a differing number of observations per hidden state, giving us variable state durations (Yu, 2010). Since not every cognitive processing step can be assumed to have the same duration, HsMMs are best suited for this analysis (Anderson et al., 2016). For our purposes, we will extract principal components from raw EEG data using principal component analysis (PCA), which is an unsupervised machine learning technique used to reduce the dimensionality of the EEG signal down to a number of principal components. These components then serve as our observation chain Y, and the underlying cognitive processing steps will be modelled by the most likely sequence of hidden states X.

To discover different processing stages, HsMM-MVPA relies on the assumption that processing stages are bookended by cognitive events that can be discovered in the EEG signal by looking for positive or negative peaks, distributed across different brain regions. This assumption is shared by two main theories explaining the generation of event-related potential (ERP); the classical theory and the synchronised oscillations theory. According to the classical theory, cognitive events generate phasic bursts of activity, where all other parts of the EEG signal are considered noise (Shah et al., 2004). Conversely, the synchronised oscillations theory proposes that these peaks result from synchronisation or phase resetting in a certain frequency band, triggered by an external event (Makeig et al., 2002). Although these theories disagree on the exact origin of the distributed peaks of activity, they agree that these peaks exist and signify cognitive events.

HsMM-MVPA searches for these positive or negative peaks in the EEG signal, distinguishing the moments of peaks and the interims in between as 'bumps' and 'flats' respectively. As an example, Figure 1 shows how the EEG signal is modelled in three different trials using HsMM-MVPA. The HsMM consists of a number of bumps that signify the onset of a new cognitive processing stage, as well as a set of gamma distributions of stage durations across trials. The algorithm first attempts to find bumps that represent the onset of a cognitive

stage and the flats that separate these bumps. The assumption exists that the EEG signal within these flats is described by sinusoidal noise around 0, but Anderson et al., 2016 have shown that HsMM-MVPA can robustly discover bumps even if this assumption does not hold. The goal of HsMM-MVPA is to identify the topology and temporal location of each bump on each trial. The method allows for variability within the duration of cognitive processes for each trial, allowing for bumps to occur at different time points per trial (but still recognising them as the same bump). The trials are analysed individually, but all trials of all participants are taken into account simultaneously. This way, inter-trial and inter-subject variability is kept to a minimum, so long as the assumption holds that every trial of every participant is under the same condition. The example HsMM-MVPA model illustrated in Figure 1 contains five bumps, each marking the start of a cognitive processing stage. There are then six stages, the first starting with stimulus presentation and the last being defined by the participant's response.

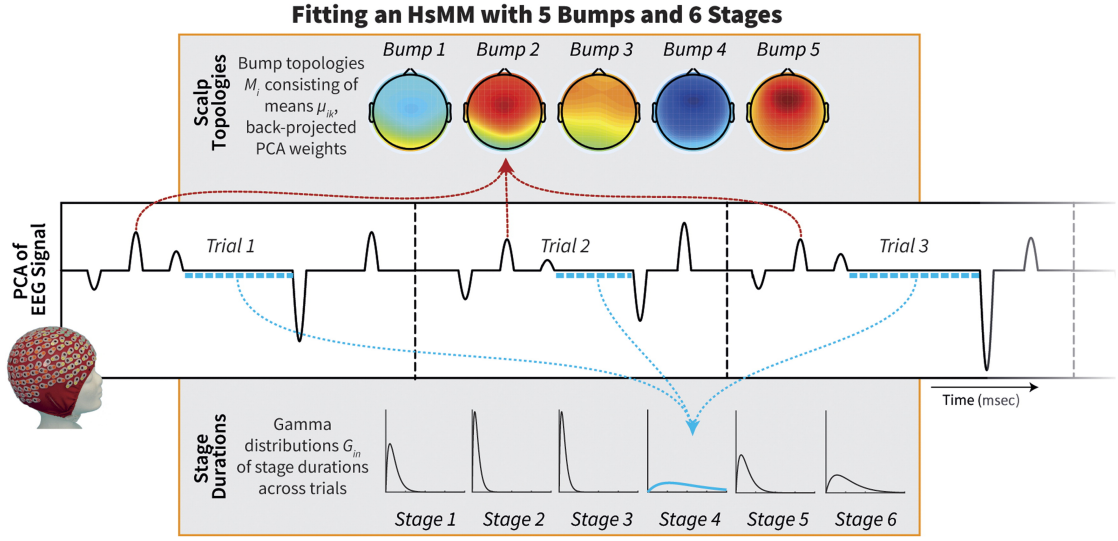


Figure 1: MVPA method applied to EEG data from three trials. The resulting model seems to contain five bumps and six stages. The location of bump 2 on each trial is signified by the red dashed arrows, and the blue dashed arrows indicate variability in the duration of stage 4 (representing the process terminated with bump 4).

A core problem in the paradigm of classification of EEG signal data is that there is a high degree of variability from subject to subject (Saha and Baumert, 2020). This means that it is difficult to classify patterns in EEG across subjects. Implementations of HsMM-MVPA as described above are able to discover good-fitting models across multiple subjects, so long as the subjects' EEG signal is collected under the same condition. This suggests that much of the variability both across trials and across subjects is accounted for by the method. The usability of the HsMM-MVPA method suggests that the across-subjects variability of EEG data lies in other areas than the number and duration of bumps and flats that the models discover. This leads to our hypothesis that the EEG signal of different subjects performing a

task under the same condition can be modelled with the same HsMM. Granted that different experimental conditions each have a model that is fitted to them, these models can be used to distinguish between these same conditions in unseen EEG data. Testing this hypothesis will be our central aim here.

For our purposes, we used EEG and behavioural data collected from individuals verifying an arithmetic problem with a given solution (Archambeau et al., unpublished). Participants were shown a multiplication problem with a solution and then asked to verify whether the given solution was correct. For instance, when a participant gets shown a multiplication problem like 6×7, they might simply retrieve the answer from memory. This is the most common way in which adults solve single-digit multiplication problems (LeFevre, Bisanz, et al., 1996). However, people can also use a procedural strategy. In the aforementioned example, the person might fail to get 6×7 from memory, but they might be able to retrieve $6 \times 6 = 36$ and then add 7. We used a self-report to assess whether someone used a procedural strategy or not (J. I. Campbell and Xue, 2001; J. I. Campbell and Timm, 2000; Grabner et al., 2009; LeFevre, Sadesky, et al., 1996; Metcalfe and Campbell, 2011).

2 Methods

In order to test our hypothesis, we constructed a classifier. First, we discover the best models for each of our two conditions. Then, we test how good these models fit our test data on a trial-by-trial basis. If performance is greater than randomness, we can conclude that inter-participant variability of the EEG signal is accounted for by HsMM-MVPA, which means that there is a demonstrable similarity between subjects' EEG data when going through the same cognitive processing steps.

2.1 Data description

We used data collected in a similar way to Archambeau et al., unpublished. The data was collected on 42 subjects between the ages of 17 and 52 ($m = 22.24$) using a BioSemi interface with 72 channels, 64 of which are data channels and 8 of which are reference channels, at a sampling rate of 2048 Hz. The study was approved by the local Ethical Review Board of the University of Amsterdam, and all participants provided informed consent. The experimental sequence is illustrated in in Figure 2. The participants were shown a single-digit multiplication sum on screen with an answer on screen and were asked to decide whether the propose solution was correct or not. Besides positive trials (where the given solution is correct) the incorrect trials are sorted into two experimental conditions: interfering solutions (I) and unrelated or non-interfering solutions (NI). With an interfering solution, the answer given is table-related to one of the operands. For example, the given interfering solution of a problem $a \times b$ could be the correct solution of $(a \pm 1) \times b$ or $a \times (b \pm 1)$. In theory, noticing that an answer is false when the given answer is table-related seems to take longer, since the activation of competing solutions produces interference during memory retrieval (J. I. Campbell and Graham, 1985; S. B. Campbell, 1995). Commutative pairs of multiplications (e.g. 6×7 and 7×6) were considered to be identical problems (J. I. Campbell, 1999; J. I. Campbell, 2005).

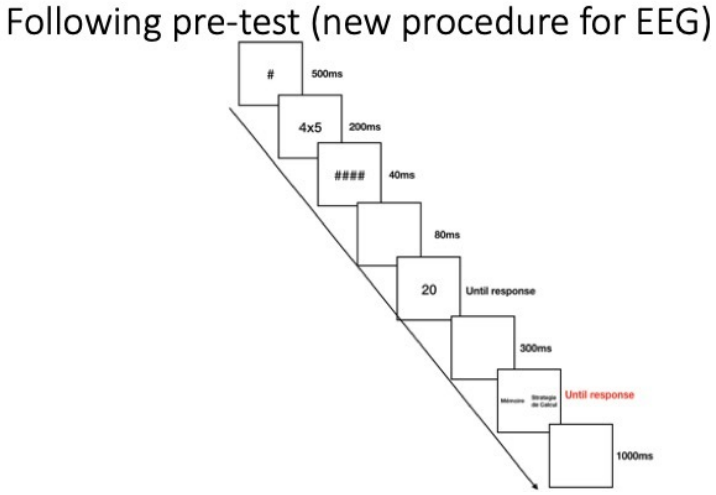


Figure 2: Multiplication experiment event sequence

Multiplication problems with a single digit as the correct solution were removed (e.g. 2×2, 3×2, etc.), as well as 9×9 as no two different interfering solutions (that are still single-digit multiplication problems) could be created for it. Half of all interfering solutions and unrelated solutions were smaller and half were larger.

Each trial started with the presentation of a fixation point of 500 ms, followed by a blank screen for 350 ms. A multiplication problem containing two operands and a solution in Arabic format (e.g. $6 \times 4 = 24$) was displayed in the centre of the screen until a response was provided. The participants were asked to indicate whether the proposed solution of the problem was correct or not, by pressing a left or right key. Participants were asked to be as fast and as accurate as possible. This screen stayed until a response is given. When a response is given, a 300 ms interval occurred, after which participants were prompted to report what strategy they used to verify the multiplication problem; "memory" or "other". Then, the next trial was initiated with an inter-trial interval of 1000ms. The task consisted of 4 blocks of 248 trials, for a total of 992 trials. There were 496 positive trials and 496 negative trials with 248 interfering solution trials and 248 unrelated solution trials. Each multiplication problem was repeated 32 times: 16 positive and 16 negative trials, half of each have interfering and the other half unrelated solutions. The multiplication problems were presented in a pseudo-randomised order, ensuring that successive problems never had the same operands. Half the participant had to press a button on the left for true, and a button on the right for false. For the other half of the participants, this mapping was reversed. The multiplication task was run on a 17-inch laptop computer, using the Psychophysics Toolbox extension (Brainard, 1997) in MATLAB (version R2013a, The Mathworks Inc., Natick, Massachusetts, USA).

2.2 Behavioural analysis

We created four subsets of the data, using the three experimental conditions positive (P), interfering (I), and not interfering (NI). We also had a subset containing all data. Subjects

32-42 were split off as a test set, making subjects 1-31 the training set. In 6.5% of trials the subject had the answer wrong. These trials were removed. Then, we removed outliers from all subsets of the data based on response times (RTs). When matching the behavioural data to the EEG data for epoching, four subjects from the training set were removed due to incorrect event numberings in the EEG data. In total, 15% of the data was removed in this step.

After cleaning our data sets, class imbalance was computed. As seen in table 1, in all subsets the vast majority of trials is labelled as 'retrieval'. These numbers will be considered as a baseline for classification accuracy. Although the standard deviation in RT in the test data subsets is much lower, HsMM-MVPA can account for variations in the temporal offset and stage durations (Anderson et al., 2016).

Condition	Percentage Retrieval	Mean RTs (ms)	SD RTs (ms)
Training data			
All	87.1%	1398	2211
P	88.5%	1312	2051
I	83.4%	1628	2101
NI	87.9%	1339	2581
Test data			
All	88.4%	1110	874
P	91.9%	1040	819
I	81.3%	1280	998
NI	84.0%	1099	837

Table 1: Overview of trials labelled "retrieval" after error and outlier removal. Standard deviation is computed within every subset, across participants.

2.3 Data preprocessing

A computational challenge with EEG data is its dimensionality and noisiness. The data we used consists of around 1-hour long recordings of 2048 samples per seconds over 64 channels. To make a model-based machine learning approach like HsMM-MVPA computationally viable, some preprocessing steps are necessary. The data was processed in MATLAB (version 2020a, The Mathworks Inc., Natick, Massachusetts, USA) using the open source EEGLab plugin (Delorme and Makeig, 2004). First, the data was high-pass filtered at 1 Hz and low-pass filtered at 40 Hz, as oscillations outside of this range are not commonly associated with brain activity (Henry, 2006). Then, the data was resampled to 512 Hz to make the following steps more computationally viable. Next, flatlines and overly noisy sections of data were removed automatically using built-in EEGlab functions, before applying Independent Component Analysis (ICA) using the FastICA algorithm (Hyvarinen, 1999). ICA is a widely used unsupervised machine learning approach that tries to decompose the EEG signal into brain- and non-brain-related components. Next, the ICLabel plugin was used to automatically flag non-brain related components from the data (Pion-Tonachini et al., 2019). These flagged

components were then removed from the data. All told, about 10% of the data was removed in this step.

2.4 Hidden semi-Markov Models

To fit the models, the data was split into a training and a test set, where the training set comprised 32 participants and the test set contained the remaining 10. The data was resampled to 100 Hz and then epoched using the behavioural data that was collected, using trial onset values that correspond to those saved in the EEG data. Four of the participants had a mismatch in the number of trials contained within their EEG data when compared to their behavioural data and were subsequently removed, leaving us with a training set of 28 participants. principal component analysis (PCA) was applied to all datasets to extract the 10 principal components from the data. In all subsets, the 10 principal components account for more than 97% of variance in the data.

We consider our bumps to have a duration of 50 ms, as this duration produces robust results even if the actual bump durations are slightly longer or shorter (Anderson et al., 2016). Each bump signifies the start of a cognitive processing stage that is continued during the subsequent flat. These flats have a mean amplitude of zero. The duration of these flats was modeled with a gamma distribution with a shape parameter of 2. The results are not sensitive to the exact choice of shape parameter, except that it simplifies the estimation of flat distributions (Anderson et al., 2016). In a model, n bumps results in $n+1$ flats (or $n+1$ processing stages), since the first stage starts with a flat when the stimulus is applied.

We constructed different HsMM-MVPA models for every subset of the data and for both cognitive strategies, so we aim to have 8 validated models by the end. Model estimation begins with a single bump model and creates models for an increasing number of bumps until a number of bumps n_{max} is reached, with n_{max} being the maximum number of 50 ms bumps that fit in the duration of the shortest epoch. During estimation, two parameters of each hidden state are obtained: (1) the amplitudes of the bumps that mark the onsets of the processing stages and (2) the scale parameter of a gamma distribution describing the stage durations (with a fixed shape parameter at 2). Data from all trials and all participants in a training subset were taken into account simultaneously. The match between the EEG data and the model was maximised using a standard expectation-maximisation (EM) algorithm (Moon, 1996).

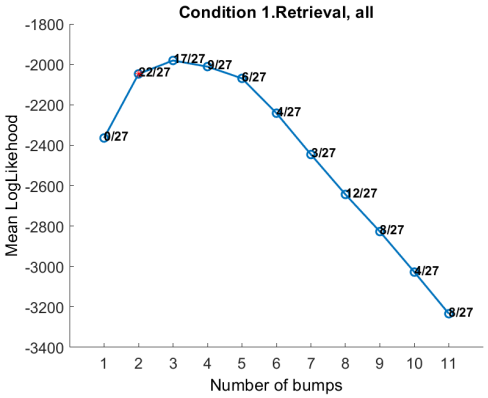
The fitting process begins by defining initial amplitudes for both the bumps and the gamma distributions for stage durations. The convergence of the EM algorithm can be sensitive to the choice of starting point, ending up in a local maximum (Wu, 1983), we used a process based on work by Zhang, Walsh, et al., 2018. Per subset, we first fit separate HsMM-MVPA models for each condition on n_{max} bumps, obtaining bump amplitudes and gamma distributions. Next, we used those parameters for models with $n_{max}-1$ bumps, iteratively leaving out each of the bumps in n_{max}, selecting the model with best fit. These bumps become the new n_{max} before the above process is repeated until only a one-bump model n_1 is left. This way, we can find all potential bump topologies while avoiding local maxima.

Since the log-likelihood tends to increase when more bumps are fitted—as there are more parameters to fit the data—we used a leave-one-out cross-validation (LOOCV) procedure to

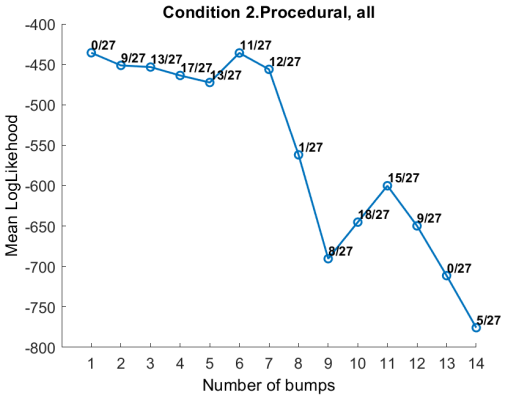
prevent overfitting. For every training subset, we estimate an HsMM-MVPA model on all participants but one and then test the fit of this model on the omitted participant. This process is repeated for all participants. Finally, we used a sign test to test for how many participants the log-likelihoods of the models with $n+1$ bumps increased compared to an n-bump model. If a model with one additional bump outperforms the previous model for a sufficiently large number of participants, we can say that the additional complexity of that model is warranted. This step is crucial for fitting models that generalise well across participants (Anderson and Fincham, 2014). For a more detailed mathematical description and code for HsMM-MVPA we refer to Anderson et al., 2016 and Berberyan et al., 2021.

2.5 Classification

After estimating the most likely parameters for both strategies, for every subset, using the method outlined above, we plotted the mean log-likelihood of all bumps, for both conditions, in all four subsets of our training data. We used these plots to select the models with the highest mean log-likelihood given the data they were trained on, per strategy and per subset. We also took into account the number of participants for whom an n-bump model improved significantly over the $n-1$-bump model. Then, we used our preprocessed test data to estimate the likelihood of every trial per subset under both models for that subset. We also estimated the likelihoods of all test trials per subset under models of different subsets to further test how well the models generalise.



(a)



(b)

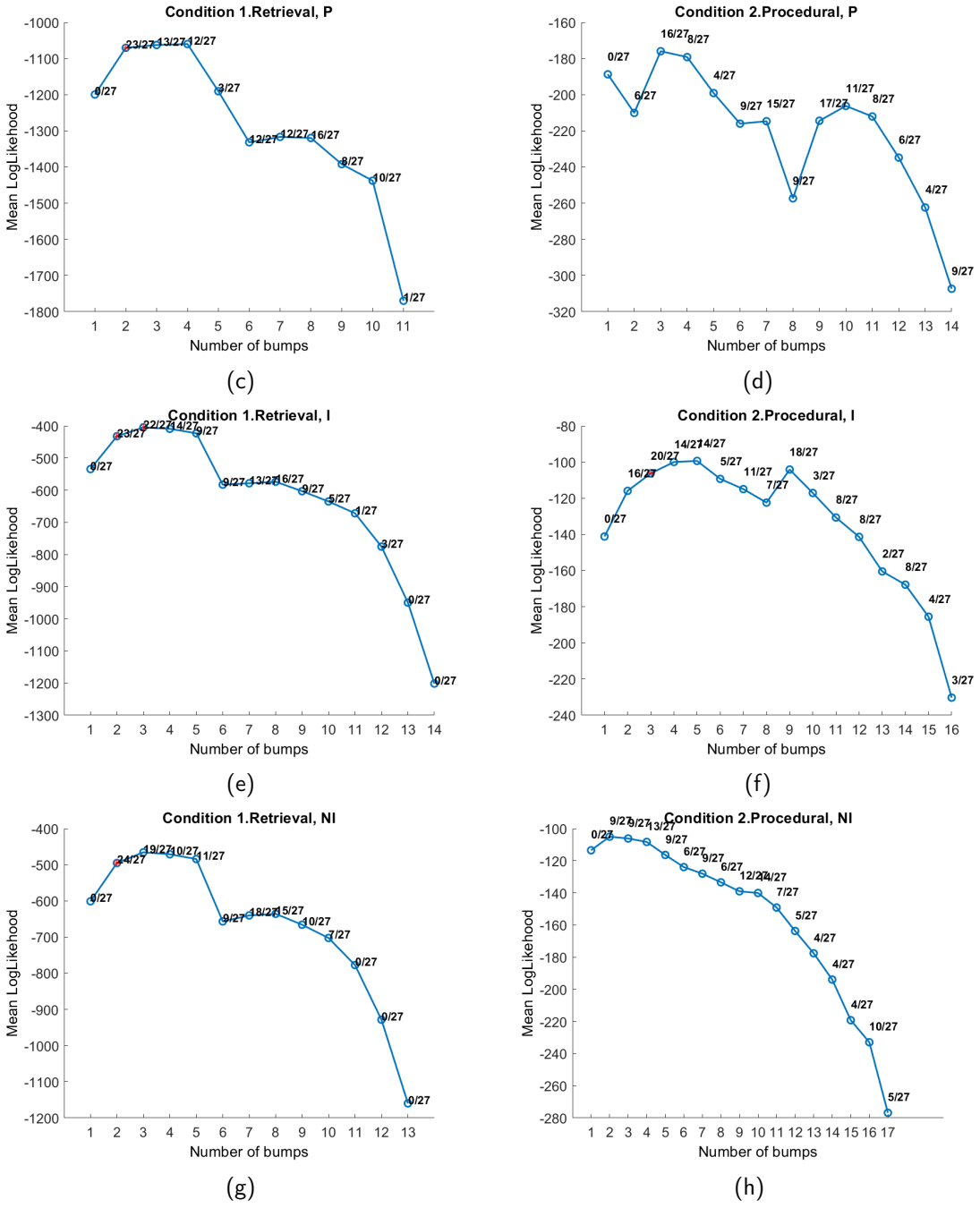


Figure 3: Model performance after estimating parameters on training data and LOOCV. The fraction next to every point indicates for how many participants the model was a significant improvement.

3 Results and Discussion

The plots of the mean log-likelihood for every number of bumps can be seen in Figure 3. Based on these results, the retrieval strategy seems to be more consistent than the procedural strategy. In every subset, the 2-bump model performs best. A 2-bump model indicates 3 processing stages, which is consistent with standard models of memory retrieval (e.g., Ratcliff, 1978). Given these results, we decided to use the retrieval-models for classification. We took our test data and computed the log-likelihood of every trial in the test data for each model. Since the actual log-likelihood values of the models can vary greatly in absolute numbers, we took the approach of creating Receiver Operating Characteristic (ROC) curves to find a log-likelihood-threshold that separated trials from both strategies with the greatest accuracy. An ROC-curve is a graphical plot used to illustrate the diagnostic ability of a binary classifier as its discrimination threshold is varied (Fawcett, 2006). It is a measure of the ratio between the true positive rate (TPR) or recall and the false positive rate (FPR). In our case, the discrimination threshold is the minimum required log-likelihood value for a single trial to be labelled as retrieval. We did this for all subsets of the test data, classifying a test data subset under the corresponding HsMM and under the other three HsMMs. The ROC curves can be seen in Figure 4.

Model trained on All			
	AUC	Max F1	Max acc
All	0.584	0.943	89.3%
P	0.612	0.974	94.9%
I	0.597	0.902	82.4%
NI	0.527	0.916	84.5%

(a)

Model trained on P			
	AUC	Max F1	Max acc
All	0.603	0.944	89.4%
P	0.653	0.974	95.0%
I	0.618	0.902	82.6%
NI	0.550	0.917	84.7%

(b)

Model trained on I			
	AUC	Max F1	Max acc
All	0.548	0.942	89.1%
P	0.557	0.974	94.9%
I	0.577	0.902	82.2%
NI	0.568	0.943	89.3%

(c)

Model trained on NI			
	AUC	Max F1	Max acc
All	0.586	0.943	89.3%
P	0.610	0.974	95.0%
I	0.598	0.902	82.4%
NI	0.529	0.916	84.7%

(d)

Table 2: Overview of model performance, listing area under the curve (AUC) and maximum accuracy and F1-scores.

All results, including area under the curve (AUC) values, can be seen in table 2. The area under an ROC curve is a quantifiable measure of the diagnostic performance of the classifier. An AUC of 0.5 denotes random performance, where an AUC of 1.0 is considered a perfect model. Overall, the ROC curves show performance better than random in most areas. The curves of models trained on all data, P, and NI show very similar curves, which suggests that the best performing models that were trained on those subsets of the training

data have a lot in common with one another. The best performing model was the model trained on P-data overall. When tested on P test data, we see an AUC of 0.653 and a classification accuracy of 95%. We also looked at the F1-scores, which is the harmonic mean of the precision and recall over the retrieval-strategy. F1-scores tend to be higher if the model is good at correctly identifying retrieval-trials, and it is less sensitive to incorrectly labelling procedural-trials as retrieval. In all models, F1-scores tend to be higher than accuracies, which means that sensitivity to the retrieval-class is better than sensitivity to the Procedural-class. The models are better at correctly recognising 'retrieval' trials than they are at correctly recognising procedural-trials as non-retrieval ones. The difference in performance between all subsets under all models is comparable to the difference in class imbalance in the test set.

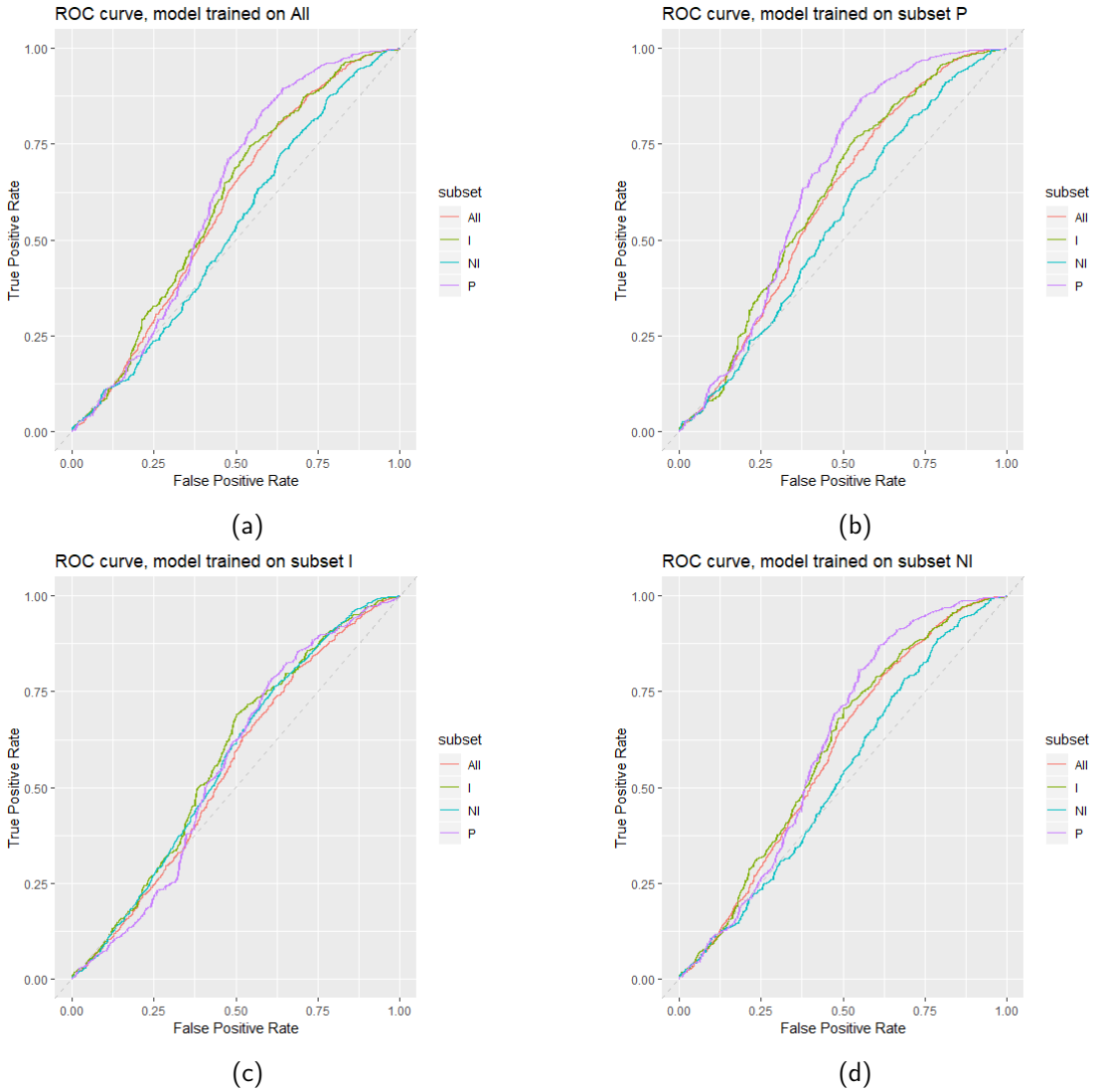


Figure 4: Receiver Operating Characteristic (ROC) curves per training data subset

4 Conclusion

In general we can say with confidence that Hidden semi-Markov Model Multivariate Pattern Analysis when applied to EEG data to detect cognitive processing stages can do so in a way that accounts for the inter-subject variability of that EEG data. In other words, an HsMM fit to a specific cognitive strategy can recognise most unseen trials of that same strategy, even when that unseen trial is from a subject whose EEG data the model has not seen at all. This implies that, when different people report using the same cognitive strategy under the same experimental condition, there is enough commonality between the cognitive processing steps that this strategy consists of, that a model-based classifier as described in this project can recognise that cognitive strategy. There is also consistency between the models fit to three of the four experimental conditions with respect to their classification on the test data subsets, but almost everywhere this consistency is proportional to the variation in the class imbalance of our data. The other side of the coin is that F1 scores everywhere tend to be higher than classification accuracy, which is closer to random performance. This means that a retrieval-model is good at recognising when a trial uses the retrieval strategy, but is less accurate when it comes to recognising that a non-retrieval trial is not retrieval.

We postulate that in our data the Procedural-strategy is less cohesive. Whenever a participant did not directly retrieve the answer from memory, they could still be employing different strategies to verify the multiplication problem that they are presented with. One can compute a multiplication problem in different ways, which might lead to different strategies. This is further supported by the fact that all well-performing Hidden semi-Markov models trained on Retrieval data had 2 bumps as the most likely amount, whereas the number of bumps varied from subset to subset under the Procedural-trials. This means that it is highly likely that participants used 3 processing stages when using the retrieval-strategy. The variance in number of bumps of the procedural models seems to suggest that what we have labelled as "procedural" actually encompasses a collection of different non-memory retrieval strategies (LeFevre, Bisanz, et al., 1996; Ashcraft, 1992). Some of these Procedural-strategies might be very similar with regards to their processing steps as the memory retrieval strategy, which explains why our classifier tends to perform the worst with regards to false positives. To complicate matters further, there is the potential of noise or biases to be introduced when using self-reports as a tool for setting our ground-truth (Kirk and Ashcraft, 2001). With respect to our data, there is a possibility that some of the procedural-trials are actually retrievals that were labelled wrongly, which might be another explanation of the difficulty that the classifier has in evaluating non-retrieval trials as such. Alternative ways of ascertaining cognitive strategies, like a mixture modelling approach (e.g. Archambeau et al., unpublished; Thevenot et al., 2007) could be considered for this purpose. There is also a discrepancy in the means and standard deviations of our response times between our training and test sets. HsMM-MVPA is able to account for variance in temporal onset and duration of processing stages (Anderson et al., 2016), so in theory this is no problem. However, this is a variable that could be corrected for in future explorations of this application of the method.

It is to be noted that the data that we used was not collected with this analysis in mind. The experimental conditions, as outlined, split the data into both two cognitive strategies and four different experimental conditions. Additionally, the cognitive strategies in the multiplication

task were essentially 'retrieval' and 'everything else', where 'everything else' might encompass cognitive strategies that are a lot like memory retrieval with regards to their processing stages as interpreted by HsMM-MVPA (Ashcraft, 1992). Given the imperfections of the data used in this analysis, we are confident that HsMM-MVPA for classification has greater promise than shown here when applied to experimental data collected for the very purpose of doing model-based classification. A classification task without a class imbalance and a greater difference between cognitive strategies might work well. Another avenue that warrants exploration is comparing a classifier that is only cross-trial to one that is cross-subject. A cross-trial classifier would have parameters fitted to some trials of a subject and then tested on unseen trials of that same subject. When we compare this to a cross-subject subject classifier, we can test whether there is a significant change in classification performance, which can give more nuanced insight into the abilities of HsMM-MVPA to estimate generalisable of processing stages. If cross-subject classification performance is comparable to within-subject classification performance, that would further support the hypothesis that HsMM-MVPA is able to account for the cross-subject variability of EEG data. With respect to the cognitive strategies investigated in this project, it could be interesting to compare a retrieval-model trained on multiplication problems to a retrieval-model trained on, for example, addition or division problems, to see whether the cognitive strategies use similar processing stages across different problem types. An analysis of this sort could tell us whether subjects using the same strategy on a different problem type actually employ the same processing stages.

In conclusion, this first investigation into using HsMM-MVPA as a tool for classification of cognitive strategies shows promise. The next step would be to investigate how far this promise can lead.

References

Anderson, J. R., Borst, J. P., Fincham, J. M., Ghuman, A. S., Tenison, C., & Zhang, Q. (2018). The common time course of memory processes revealed. *Psychological science*, *29*(9), 1463–1474.

Anderson, J. R., & Fincham, J. M. (2014). Extending problem-solving procedures through reflection. *Cognitive psychology*, *74*, 1–34.

Anderson, J. R., Lee, H. S., & Fincham, J. M. (2014). Discovering the structure of mathematical problem solving. *NeuroImage*, *97*, 163–177.

Anderson, J. R., Zhang, Q., Borst, J. P., & Walsh, M. M. (2016). The discovery of processing stages: Extension of sternberg's method. *Psychological review*, *123*(5), 481.

Archambeau, K., Molenaar, D., Forstmann, B., Noël, M.-P., Gevers, W., & Van Maanen, L. (unpublished). Age-related differences in the resolution of interference in simple arithmetic depends on the strategy used.

Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition*, *44*(1-2), 75–106.

Berberyan, H. S., van Maanen, L., van Rijn, H., & Borst, J. (2021). Eeg-based identification of evidence accumulation stages in decision-making. *Journal of Cognitive Neuroscience*, *33*(3), 510–527.

Borst, J. P., & Anderson, J. R. (2015). The discovery of processing stages: Analyzing eeg data with hidden semi-markov models. *NeuroImage*, *108*, 60–73.

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

Campbell, J. I. (1999). Division by multiplication. *Memory & Cognition*, *27*(5), 791–802.

Campbell, J. I. (2005). *Handbook of mathematical cognition*. Psychology Press.

Campbell, J. I., & Graham, D. J. (1985). Mental multiplication skill: Structure, process, and acquisition. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *39*(2), 338.

Campbell, J. I., & Timm, J. C. (2000). Adults' strategy choices for simple addition: Effects of retrieval interference. *Psychonomic Bulletin & Review*, *7*(4), 692–699.

Campbell, J. I., & Xue, Q. (2001). Cognitive arithmetic across cultures. *Journal of experimental psychology: General*, *130*(2), 299.

Campbell, S. B. (1995). Behavior problems in preschool children: A review of recent research. *Journal of child Psychology and Psychiatry*, *36*(1), 113–149.

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.

Donders, F. C. (1868). Over de snelheid van psychische processen. *Onderzoekingen gedaan in het Physiologisch Laboratorium der Utrechtsche Hoogeschool (1968–1869)*, *2*, 92–120.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Grabner, R. H., Ansari, D., Koschutnig, K., Reishofer, G., Ebner, F., & Neuper, C. (2009). To retrieve or to calculate? left angular gyrus mediates the retrieval of arithmetic facts during problem solving. *Neuropsychologia*, *47*(2), 604–608.

Harding, B., Goulet, M.-A., Jolin, S., Tremblay, C., Villeneuve, S.-P., & Durand, G. (2016). Systems factorial technology explained to humans. *Tutorials in Quantitative Methods for Psychology*, *12*(1), 39–56.

Henry, J. C. (2006). Electroencephalography: Basic principles, clinical applications, and related fields. *Neurology*, *67*(11), 2092–2092.

Henson, R. (2011). How to discover modules in mind and brain: The curse of nonlinearity, and blessing of neuroimaging. a comment on sternberg (2011). *Cognitive neuropsychology*, *28*(3-4), 209–223.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, *10*(3), 626–634.

Imani, E., Harati, A., Pourreza, H., & Goudarzi, M. M. (2020). Brain-behaviour relationships in the perceptual decision-making process through cognitive processing stages. *bioRxiv*.

Kirk, E. P., & Ashcraft, M. H. (2001). Telling stories: The perils and promise of using verbal reports to study math strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 157.

LeFevre, J.-A., Bisanz, J., Daley, K. E., Buffone, L., Greenham, S. L., & Sadesky, G. S. (1996). Multiple routes to solution of single-digit multiplication problems. *Journal of Experimental Psychology: General*, *125*(3), 284.

LeFevre, J.-A., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 216.

Makeig, S., Westerfield, M., Jung, T.-P., Enghoff, S., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2002). Dynamic brain sources of visual evoked responses. *Science*, *295*(5555), 690–694.

Metcalfe, A. W., & Campbell, J. I. (2011). Adults' strategies for simple addition and multiplication: Verbal self-reports and the operand recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 661.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, *13*(6), 47–60.

Pinheiro-Chagas, P., Piazza, M., & Dehaene, S. (2019). Decoding the processing stages of mental arithmetic with magnetoencephalography. *cortex*, *114*, 124–139.

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). Iclabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197.

Portoles, O., Borst, J. P., & van Vugt, M. K. (2018). Characterizing synchrony patterns across cognitive task stages of associative recognition memory. *European Journal of Neuroscience*, *48*(8), 2759–2769.

Posner, M. I. (2005). Timing the brain: Mental chronometry as a tool in neuroscience. *PLoS Biol*, *3*(2), e51.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, *85*(2), 59.

Saha, S., & Baumert, M. (2020). Intra- and inter-subject variability in eeg-based sensorimotor brain computer interface: A review. *Frontiers in Computational Neuroscience*, *13*, 87. https://doi.org/10.3389/fncom.2019.00087

Shah, A. S., Bressler, S. L., Knuth, K. H., Ding, M., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2004). Neural dynamics and the fundamental mechanisms of event-related brain potentials. *Cerebral cortex*, *14*(5), 476–483.

Sternberg, S. (1969). The discovery of processing stages: Extensions of donders' method. *Acta psychologica*, *30*, 276–315.

Thevenot, C., Fanget, M., & Fayol, M. (2007). Retrieval or nonretrieval strategies in mental arithmetic? an operand recognition paradigm. *Memory & Cognition*, *35*(6), 1344–1352.

Visser, I., Raijmakers, M. E., & van der Maas, H. L. (2009). Hidden markov models for individual time series. *Dynamic process methodology in the social and developmental sciences* (pp. 269–289). Springer.

Walsh, M. M., Gunzelmann, G., & Anderson, J. R. (2017). Relationship of p3b single-trial latencies and response times in one, two, and three-stimulus oddball tasks. *Biological psychology*, *123*, 47–61.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, 95–103.

Yu, S.-Z. (2010). Hidden semi-markov models. *Artificial intelligence*, *174*(2), 215–243.

Zhang, Q., Borst, J. P., Kass, R. E., & Anderson, J. R. (2017). *Inter-subject alignment of meg datasets in a common representational space* (tech. rep.). Wiley Online Library.

Zhang, Q., van Vugt, M., Borst, J. P., & Anderson, J. R. (2018). Mapping working memory retrieval in space and in time: A combined electroencephalography and electrocorticography approach. *NeuroImage*, *174*, 472–484.

Zhang, Q., Walsh, M. M., & Anderson, J. R. (2017). The effects of probe similarity on retrieval and comparison processes in associative recognition. *Journal of Cognitive Neuroscience*, *29*(2), 352–367.

Zhang, Q., Walsh, M. M., & Anderson, J. R. (2018). The impact of inserting an additional mental process. *Computational Brain & Behavior*, *1*(1), 22–35.

Zylberberg, A., Dehaene, S., Roelfsema, P. R., & Sigman, M. (2011). The human turing machine: A neural framework for mental programs. *Trends in cognitive sciences*, *15*(7), 293–300.