Consistency in Extending Problem-solving Procedures Indicates Expertise

Qiong Zhang
DAP Committee: John Anderson, Rob Kass, Roy Maxion

December 18, 2014

Background. Transfer of learning, which is the extent that one is able to apply prior experience and knowledge in a new but similar situation, has been studied extensively since the beginning of the last century. Recent brain imaging studies have provided new insight into how students are able to extend their previous problem solving skills to new but similar problems.

Aim. We wish to determine what contributes to individual differences in the success at transfer of learning, which is measured by consistency – the extent of agreement among activated brain-image voxels from one brain region of interest (ROI) to another.

Data. 75 participants were familiarized with solving a set of mathematical problems before being put into an fMRI scanner, where they were challenged to solve modified versions of the same set of familiar problems. Brain activations were recorded as a time series of volumetric structures of 8365 voxels for each fMRI scan.

Methods. A hidden semi-Markov Model identified the sequential structure of thought when solving the problems. Brain consistencies at different levels of brain analysis were characterized.

Results. Analyzing the patterns of brain activity over the sequence of states identified by the model, we observed that there was less brain consistency (agreement) among weaker-performing subjects than among better subjects in an ANOVA analysis ($F(1, 73) = 20.40, p < .0005$). In particular, early consistency was predictive of overall performance in the experiment. In addition to looking at activity across all brain voxels and in pre-defined brain regions, a data-driven approach over the whole brain was carried out as an alternative analysis that verified the same observation.

Conclusion. Brain consistency is not just an indicator of subject performance during transfer of learning, but also a predictor of overall performance using only a few observations at the early stage of the experiment. The observed effect of consistency is due to perceived similarity between the new problems and the trained problems during the transfer. The results hold for alternative measures of brain consistency and for different levels of brain analysis.

1 Introduction

There has been interest in understanding how the human mind is able to handle novel instances of a familiar problem. While sometimes the learners are only skilled at what is

exactly being taught, in many cases they are expected to transfer what they have learned to new situations, in what we call transfer of learning. Transfer of learning is the application of knowledge and experience gained in one setting to another setting [1]. It is the fundamental assumption in education that what is learned will apply in similar but possibly different situations [2]. A specific setting that has raised considerable interest is the transfer of learning of mathematical problem solving.

For decades, different methods have been used to shed light on the underlying thought processes during complex mathematical problem solving. Such efforts include protocol analysis that relies on verbal evidence and multi-voxel pattern analysis (MVPA) techniques that recognize the representational structures in the brain. As functional Magnetic Resonance Imaging (fMRI) is becoming a powerful instrument to collect vast quantities of data on brain activity, a new procedure has emerged that combines MVPA and Hidden Markov Model (HMM) algorithms [3] to better integrate the temporal patterns in the brain [4]. This method is particularly effective in mathematical problem solving, where there is a rich mixture of perceptual, cognitive and motor activities with distinct temporal characteristics. In looking into the underlying thought processes of such a task, a hidden semi-Markov model is used in our study to identify sequential structure and durations of problem solving.

In recent years, brain imaging studies have informed us about the neural basis and mechanisms underlying transfer of learning [5] [6]. However, there remains an open question: what are the sources of individual differences in successful transfer? This question is the focus of our study. In our experiments, participants were trained in a mathematical problem-solving task before they were scanned. In the fMRI sections of the experiments, they encountered both Regular problems that were like those they had solved, as well as novel Exception problems that required participants to devise modifications or partial replacements to their learned procedure.

2 Problem being solved

This is a exploratory analysis in which we attempt to find an answer to the question: what contributes to individual differences in transfer-of-learning performance?

3 Approach

Early evidence in behavior experiments demonstrated that the extent to which subjects are able to relate the current task to the recalled learned task contributes to success in transfer tasks [7]. There is a tendency to apply consistent strategies once the learned task is recalled. We ask the question in our study: is there similar evidence at the brain level that can be better quantified? Does consistency give us any information as to how well someone is performing the transfer task? We start our investigation by characterizing what are considered as brain consistencies. To be intuitively comparable to behavior experiments, brain consistency is defined as the agreement in brain activation patterns amongst different problems in an experiment for a given subject. These brain consistencies are first explored on a set of pre-defined regions of interest (ROIs), then tested on a set of predictive brain voxels, and lastly verified at the whole-brain level.

4 Methods

4.1 Materials

We use a pyramid experiment – a type of mathematical problem solving – which uses a dollar symbol as the operator, e.g., 4\$3 = X. Here 4 is the base, which is also the first term in an additive sequence; 3 is the height, which indicates the number of terms to add in a descending manner, e.g., 4\$3 = 4 + 3 + 2. This is an example of a "Regular" problem. There are two types of "Exception" problems. One type has unusual numbers, being either negative, e.g., 4\$-3=X or large, e.g., 208\$3 = X. The other type requires an unusual algorithm having the unknown X on the left-hand side of the equation, e.g., X\$4 = 30. The procedural details are described in the original empirical report [8]. Here are some worked examples:

- 6 \$ 5 = X → 6 + 5 + 4 + 3 + 2 = 20
- -9 \$ 4 = X → -9 + -10 + -11 + -12 = -42
- X \$ 4 = X → 2 + 1 + 0 + -1 = 2

4.2 Subjects

fMRI data was collected from 40 adults recruited at Carnegie Mellon University including undergraduate students and graduate students (ages 19-35) and 35 children recruited from local schools in Pittsburgh (ages 12-14).

4.3 Instructions to the subjects

```
There is a notation for writing repeated addition where each term added
is one less than the previous:

For instance 4 + 3 + 2 is written as 4 $ 3

Since 4 + 3 + 2 = 9 we would evaluate 4 $ 3 as 9 and write 4 $ 3 = 9
The parts of 4 $ 3 are given names:

4 is the base and reflects the number you start with
3 is the height and reflects the total number of items you add,
including the base
4 $ 3 is called a pyramid

In this session, you will solve a series of these problems. For
example, if you see 4 $ 3 = X, type 9 on the keypad and press enter.
```

4.4 Procedure / Data Acquisition

Participants practiced solving a large number of Regular problems on a prior day outside of the scanner, with the second day tested in an fMRI scanner with a mixture of Exception problems and Regular problems. Each subject solved six blocks of problems with each block consisted of 2 Regular problems and 9 Exception problems.

Images were acquired using gradient echo-echo planar image (EPI) acquisition on a 3T Verio, then motion corrected and co-registered. The BOLD response recorded from fMRI is de-convolved with a hemodynamic response function to produce an estimate of the underlying activity signal using a Wiener filter [9]. The hemodynamic function is the difference of two gamma bases [10].

4.5 Analysis

4.5.1 Dimensionality Reduction

Multi-voxel pattern matching (MVPA) was carried out as a step of dimensionality reduction, as well as accommodating variations in anatomy over participants. Voxels are aggregated into 2x2 larger regions, with those voxels showing extreme values removed. The Blood Oxygenation Level Dependent signal (BOLD) response is calculated as the percent change from a linear baseline defined from the first scan. This is de-convolved with a hemodynamic response function to produce an estimate of the underlying activity signal. To further reduce the dimensionality of 8365 voxels, a principal component analysis was performed with 67% of the variance captured by the first 20 components that we eventually worked with. We examined the BOLD brain activations of the 14 key regions of interest (ROIs) averaged over the left side and the right side of the brain. The 14 ROIs are contained in the Appendix.

4.5.2 Discovering Mental States

Hidden Markov models (HMM) have been successfully used in modeling and analyzing complex behavioral and neurophysiological data [11]. They make inference of unobserved parameters possible while taking into account the probabilistic nature of behavior and brain activity. We conceive the participants as going through a sequence of mental states in the pyramid experiments. The discrete mental states are hidden in the sense that we only observe the brain activity, not the mental states themselves. The effectiveness of HMMs in modeling such an experiment has already been demonstrated in a previous study [8]. The real interest now is to discover the sequential structure within the problem-solving stage using the HMM without the ground truth. The optimal number of states in problem solving has been identified previously by applying a procedure of leave-one-out cross-validation (LOOCV). Parameters are estimated by maximizing the likelihood for all but one of the participants, and then used to calculate the likelihood of the data for the remaining participant. This process is rotated through k subjects. A model with n+1 states is justified over an n-state model if it increases the likelihood of the data for a significant number of subjects in this LOOCV. Four states were identified as the optimal number in the previous study with the same experiment. They are characterized as an encoding stage, a planning stage, a solving stage and a responding stage, given their brain signatures [8].

Model specification. A specific extension of HMM, a hidden semi-Markov model (HSMM), is used to model explicitly the state duration as a gamma distribution. Such

a parametric distribution is not only more realistic and widely used in modeling response latencies [12] but it also significantly reduces the number of parameters to estimate for the model. The state duration is discretized to the nearest scan. The probability of spending m scans in state i, given the length of each scan being 2 seconds, is as below:

$$G(m; v_i, a_i) = \int_{2m-1}^{2m+1} g(t; v_i, a_i)\, \mathrm{d}t$$

The fMRI activity considered in the model are the first 20 components obtained from the Principal Component Analysis over all scans in the experiment. They are further normalized to have mean 0 and standard deviation 1. The brain activity of the kth PCA component for each state i is modeled as a normal distribution $N(x; \mu_{ik}, 1)$. The probability of observing a set of PCA components $F_j = \{f_{j1}, f_{j2}, .., f_{j20}\}$ for a particular scan j, at state i, is calculated as below:

$$P(F_j; M_i) = \prod_{k=1}^{20} N(f_{jk}; \mu_{ik}, 1)$$

We use only the first 20 PCA components, and we assume parametric distribution for the same purpose of reducing the number of parameters. Let λ stand for the complete set of model parameters. Other than the parameters that define gamma distribution G of state duration and normal distribution P of observation distribution, λ also includes the transition probabilities $\{a_{mn}\}$. We conceive the participants as going through a sequence of mental states - Encode, Plan, Solve and Respond. Thus it is a left-right HSMM that adapts a linear structure of 4 mental states plus a rest state. State skipping is allowed to account for trials with very short length. The original scans have been inserted with rest scans in between two trials to make sure that the model will be forced into the rest state during rest scans and restart from the first state when a new trial starts.

Implementation. The implementation that we adapt in our HSMM is a redefined forward-backward (FB) algorithm [13] that can avoid the underflow problem in practical applications without being more complex than the most efficient FB algorithm proposed by Yu [14]. In the FB algorithm, model parameters are re-estimated at each round through maximum *a posteriori* (MAP) until the likelihood of the observations converges to a certain value.

Trial-alignment by states. Modeling with an HSMM not only uncovers sequential stages of problem solving, but also serves as an effective way to align fMRI data on a trial-by-trial basis, as was pointed out in an earlier application of HSMM on EEG data studying associative recognition [15]. Cognitive processes can be highly variable in their durations from trial to trial and from person to person; it is challenging to have them aligned properly before any further analysis. Conventionally, trials would be aligned to a fixed time point, either the stimulus or the response. However, because different processes will occur at the same time on different trials, the average signal will be uninformative [16] [17]. In our analysis, after fitting all trials to the HSMM, state occupancies for each scan $p(q_t = i)$ can be obtained. They are the probabilities of each scan belonging to each of the four states, where q_t is the state at time t, and i = 1,2,3,4. As is illustrated in Figure 1, brain activations for each scan are converted to brain activations for each state through a weighted sum by the state occupancies. These state-specific brain activations for each brain voxel will be used throughout the analysis of our study. Detailed procedures for estimating parameters of the HSMM are in the Appendix.

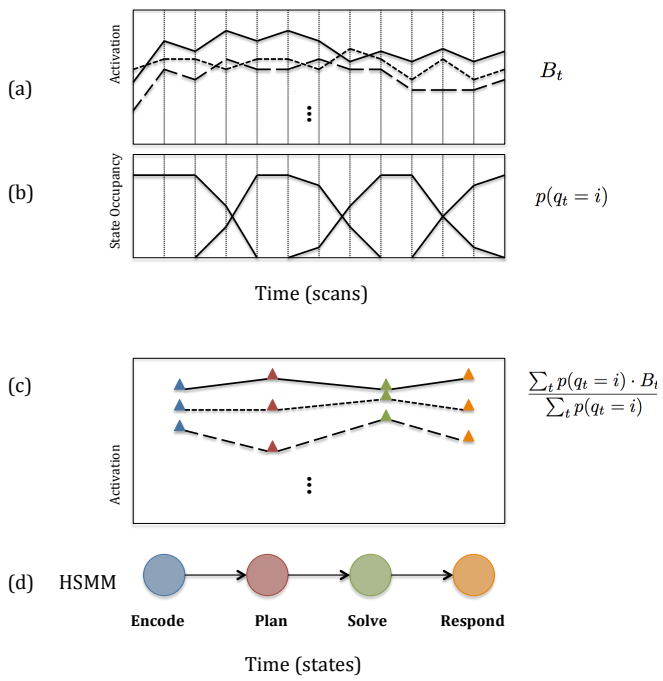

Figure 1: State-specific brain activations for each of the 4 states in a single trial for different brain regions, i = 1,2,3,4. (a) shows the brain activations for each scan for representative brain regions in one trial. (b) shows the state occupancy estimated from the HSMM of that trial. (c) shows the state-specific brain signatures calculated from weighted sums for each region in (a). (d) shows the linear structure of the HSMM model. Each vertical line in (a) and (b) represents an fMRI scan.

4.6 Design

This section describes the three experiments conducted in this study:

- Experiment 1 is designed to test whether the better subjects have a higher brain consistency, measured over 14 selected brain regions, than the weaker subjects.
- Experiment 2 is designed to test whether the better subjects have a higher brain consistency, measured over pre-selected brain voxels, than the weaker subjects.
- Experiment 3 is designed to test whether the better subjects have a higher brain consistency, measured over the entire brain volume, than the weaker subjects.

4.6.1 Experiment 1: the pre-defined brain regions

A set of pre-defined brain regions of interest (ROI) have been observed to play an important role in the complex problem solving in our laboratory. We investigated how the consistency in a participant's activation predicted the proportion of problems that participants got correct on this set of 14 ROIs. This consistency can be understood from two perspectives. It can be either measured as the correlation between the subject mean and the population mean (between-subjects consistency) or measured within each subject as the averaged correlation between every pair of problems (within-subject consistency) across the 14 ROIs.

An estimate of brain activations is obtained for the four states identified by the HSMM (encoding, planning, solving, and responding). Correlation can be computed for each state. Analysis is done on either the first 10 correctly solved problems or the last 10 correctly solved problems for each subject, in order to compare the difference between the early stage of experiment and the late stage of the experiment. To further quantify the individual differences, a second level of correlation is carried out between the within-subject brain consistency (average correlation) and the subject overall performance. The larger the value is, the more correlated the measure of consistency with the subject performance. To see how well we could use this within-subject brain consistency for prediction we used a Leave-one-subject-out cross validation (LOOCV) procedure. In particular, the performance of an unseen subject is predicted by weights trained from the rest of the subjects using multiple regression analysis with Least Squares Fitting. The four independent variables (predictors) here are the measures of consistency among the first 10 correctly solved problems for each of the four states. The dependent variable (what we are predicting) is the subject performance.

4.6.2 Experiment 2: the selected brain voxels

Previously, brain consistency was measured as correlation across the pre-defined 14 ROIs. Rather than focusing on these predefined regions, we can apply the within-subject consistency analysis to the most predictive brain regions obtained from an initial exploratory step. This is done by only considering the brain voxels whose average activation on the first 10 correctly solved problems significantly correlated with the subject performance ($r = .2272, p < 0.05$, two-tailed). To quantitatively examine how well the measure of consistency can predict the subject performance, a similar procedure of leave-one-subject-out cross validation (LOOCV) is used. For each state of each subject we calculate the within-subject brain consistency – the mean correlation of activations in the selected regions over every pair of problems during the early stage of the experiment. Then we regress the 4 mean correlations of the 74 subjects against their overall performance. This regression could then be used to predict the performance of the 75th subject, given the 4 brain consistencies corresponding to each state. To avoid selecting the seemingly most predicted brain voxels by chance, the brain voxels are selected from the 74 subjects.

4.6.3 Experiment 3: whole brain parcellation

In this section, a different measure of brain consistency at the level of the whole brain is proposed that includes the entire brain volume without a pre-selection procedure on voxels or regions. Analysis based on a few regions of interest (ROIs) could potentially omit certain information outside the regions. Definition of brain atlases partly address this issue by providing a set of ROIs that cover the entire brain volume, the construction of which is heavily based on an ontology of brain structures given current knowledge [18]. Besides the concern that there exist multiple inconsistent brain atlases, it is also questionable whether a given atlas may generalize well over the entire population [19].

To better account for individual differences, a data-driven approach – brain parcellation – is often used before further analysis where functional homogeneous brain voxels are grouped together. It is mostly used in the literature as a crucial step for data reduction, as one would then only look at brain activation averaged over regions instead of those of all the brain voxels [20]. It is also useful in deriving representative nodes before constructing a brain network in carrying out functional connectivity analysis [21].

In this study, however, we propose the use of data-driven brain parcellation as a method for comparing the brain consistencies across different problems. The consistency of two brain patterns here does not depend directly on individual response of voxel activations, but on how similarly brain voxels are grouped together when spectral clustering is applied. This is similar to an earlier approach where consistency of brain networks was measured by similarity of identified key nodes with the Jaccard Index [22], but different in the sense that there are multiple groups generated in the process of brain parcellation instead of only two groups (key-nodes and non key-nodes) which will give more detailed characterization of local brain information.

In our study, for each of the 8365 brain voxels, there are 4 brain activations for each problem associated with states of Encode, Plan, Solve and Respond. Spectral clustering is able to group the brain voxels with similar responses to the four states, while at the same time taking into account a spatial constraint where only adjacent brain voxels could be grouped together.

The Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) both measure how consistent two clusterings are to each other [23]. They are used to evaluate the consistency between two brain parcellations that correspond to solving two different problems. An estimation of brain consistency for any selected 10 problems is thus the averaged ARI/AMI values across every pair of the problems. Detailed procedures of obtaining both the two measures and the spectral clustering are described in the Appendix. We calculate brain consistencies measured from each subject and then correlate them with the subject overall performance.

5 Results and Discussions

5.1 Result 1: The Pre-defined Brain Regions

Experiment 1 was designed to test whether the better subjects have a higher brain consistency, measured over 14 selected brain regions, than the weaker subjects. A number of phenomena were observed.

5.1.1 Observation 1: the group of better subjects exhibits a higher level of within-subject consistency than the group of weaker subjects, especially at the early stage

The 75 subjects can be divided into two groups based on their performance – better subjects (38) and weaker subjects (37). The problems correlated can be divided into the early stage (the first 10) and the late stage (the last 10). As illustrated in Figure 2, at the early stage of the problem solving, the group of better subjects exhibits higher level of within-subject consistency than the group of weaker subjects. An analysis of variance (ANOVA) was performed on these average correlations where the factors were group (better vs. weaker), period (first 10 versus last 10), and state (encode, plan, solve, respond). There are significant effects of all three factors (group: $F(1, 73) = 20.40, p < .0005$), period: $F(1, 73) = 20.03, p < .0001$), and state: $F(3, 219) = 45.51, p < .0001$). There was also a significant interaction between group and period ($F(1, 73) = 5.06, p < .05$) such that the difference between better and weaker subjects decreases from (.611 versus .432) for the first 10 problems to a smaller difference (.490 versus .392) for the last 10 problems as what is illustrated in Figure 2 .

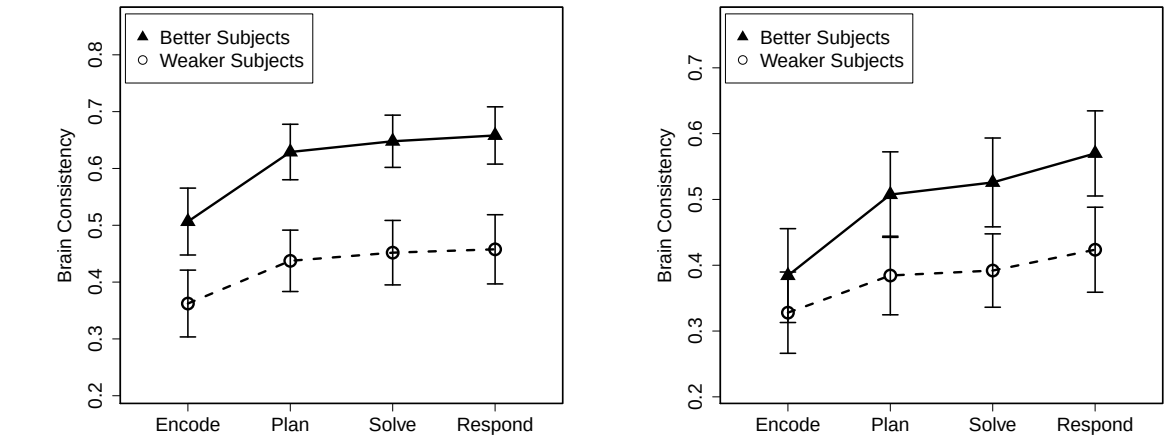

Figure 2: Brain consistency (within-subject) for the better subjects and the weaker subjects. (a) shows that during the first 10 correctly solved problems the group of better subjects exhibits a higher level of within-subject consistency than the group of weaker subjects. (b) shows that during the last 10 correctly solved problems the effect is weaker. Error bars show the 95% confidence interval of the population means.

To further quantify the relation at the level of individual difference, a second level of correlation is carried out between within-subject brain consistency (average correlation) and subject overall performance. The larger the values are, the more correlated the measure of consistency with the subject performance. Brain activations in this analysis are obtained with respect to each of the four states - Encode, Plan, Solve and Respond. An Overall correlation is also obtained through collapsing the 4 states to a single vector of 56 (14x4) regions. In our analysis, Pearson's Correlation Coefficient of larger than 0.2272 is considered significant ($p < 0.05$) under two-tailed probabilities with a sample of size of 75 (df = 73). As is shown in Table 1, there is significant correlation between within-subject brain consistency and subject overall performance for both overall and for each of the 4 states. This effect is particular strong in the first 10 correctly solved problems than the last 10 correctly solved problems.

HSMM states	Encode	Plan	Solve	Respond	Overall
First 10	**0.455**	**0.607**	**0.578**	**0.504**	**0.558**
Last 10	0.164	**0.336**	**0.364**	**0.360**	**0.365**

Table 1: Pearson correlations between subject performance and the *within-subject* brain consistency among either the first 10 correctly solved trials or the last 10 correctly solved trials, for each of the four states: Encode, Plan, Solve and Respond. For example, the 0.455 indicates that there is a significant correlation between subject performance and the within-subject brain consistency among the first 10 correctly solved trials for the encode state. An overall correlation is also obtained through collapsing the 4 states to a single vector of 56 (14x4) regions. Significant ($p < 0.05$, two-tailed) correlations are in bold.

5.1.2 Observation 2: The group of better subjects exhibits a higher level of between-subjects consistency than the group of weaker subjects

There is also significant correlation between between-subjects consistency and subject overall performance as shown in Table 2 analyzed in a similar manner as that of the within-subject consistency. An analysis of variance (ANOVA) was performed on the correlations between the subject mean and the global mean where the factors were group (better vs. weaker), period (first 10 versus last 10), and state (encode, plan, solve, respond). There is significant effect of the factor (group: $F(1, 73) = 7.62, p < .01$).

HSMM states	Encode	Plan	Solve	Respond	Overall
First 10	**0.241**	**0.316**	**0.330**	**0.336**	**0.335**
Last 10	**0.269**	**0.373**	**0.352**	0.177	**0.296**

Table 2: Pearson correlations between subject performance and *between-subjects* brain consistency among either the first 10 correctly solved trials or the last 10 correctly solved trials, for each of the four states: Encode, Plan, Solve and Respond. For example, the 0.241 indicates that there is a significant correlation between subject performance and the between-subjects brain consistency among the first 10 correctly solved trials for the encode state. An overall correlation is also obtained through collapsing the 4 states to a single vector of 56 (14x4) regions. Significant ($p < 0.05$, two-tailed) correlations are in bold.

Figure 3 illustrates the above results as two point clouds. The two clouds represent the first 10 problems of a better subject and of a weaker subject. Taking the population mean as an estimation of how one should behave, the cloud of points that is located closer to the population mean corresponds to the better subject with higher between-subjects consistency. This observation relates to a previous study of the Space Fortress task where a global definition of mental stages obtained from all the subjects can predict better the mental stage of a particular scan for individual subjects [24].

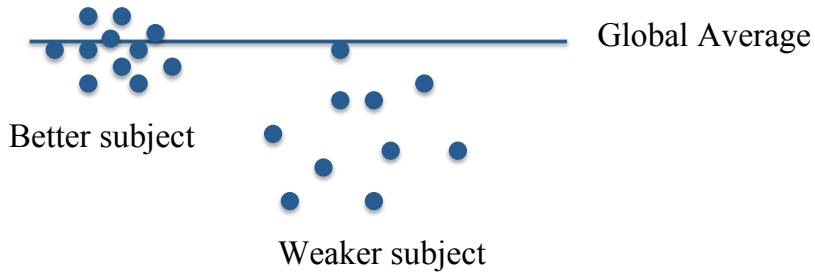

Figure 3: Illustration of the effects of within-subject consistency and between-subjects consistency. Each point represents a problem, and each cloud of points represents a subject. The better subject (left) has a higher within-subject consistency with a small dispersion of points, and also has a higher between-subjects consistency being closer to the global average.

The dispersion of the points within each subject represents how similarly and consistently the subject responds to the first 10 problems. The cloud of points with a smaller dispersion

corresponds to the better subject with higher within-subject consistency. It is interesting how the derivation of one's brain response to the global average would say about how one performs. However we will focus for the rest of the study on the examination of the within-subject consistency whose effect is stronger, and argue that this effect arises specifically from the transfer task.

5.1.3 Observation 3: Early consistency is predictive of the overall performance

To see how well we could use the within-subject brain consistency for prediction we used a leave-one-subject-out cross validation (LOOCV) analysis. As is observed in Figure 4, there is a considerable match between the predicted performance using LOOCV and the actual accuracy of the subjects, with the correlation of the two being 0.561, and the mean squared error (MSE) being 0.0399, which is also the leave-one-out cross validation estimate of the predictive risk. Thus it can be concluded that the within-subject brain consistency among the first 10 correctly solved problems is not only an indicator but also an effective predictor of the performance of an unseen subject.

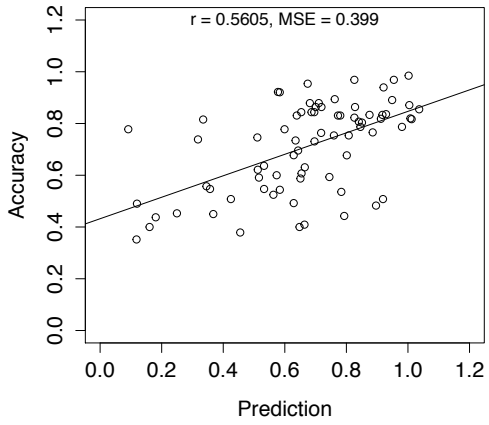

Figure 4: Leave-one-subject-out cross validation (LOOCV) performance prediction of 75 subjects compared with the actual subject performance, using the weights of multiple regression analysis obtained from the rest of the subjects. A total of 4 predictors are 4 within-subject brain consistencies measured across 14 ROIs among the first 10 correctly solve problems for each of the 4 states. (r = .5605, MSE = .0.399).

The same analysis was also carried out using the late brain consistency instead of the early consistency, which gives 0.361 as the correlation between the predicted performance and the actual accuracy, and 0.0786 as MSE. To statistically verify if a linear regression model using 8 predictors, including both early consistency and late consistency, will produce significantly better prediction than a linear regression model with only 4 predictors from late consistency, an F-test is applied which gives $p = 0.00018(< 0.05)$. Meantime, if we compare a model using 8 predictors with a model using only 4 predictors from early consistency, the F-test gives $p = 0.517(> 0.05)$. It can be concluded that including the early consistency will improve the prediction using the late consistency, but not vice versa.

5.1.4 Observation 4: Individual difference is due to the perceived similarity between the trained (Regular) problems and the new (Exception) problems

Extending the problem solving procedure from a familiar problem to a new one is a challenging task. Subjects were trained on the Regular problems during pyramid experiments before they were exposed to the Exception problems during the fMRI scanning. How successfully they would be able to handle a similar but different situation after mastering the previous task is within the scope of studying transfer of learning.

At the beginning of the last century, Thorndike and Woodworth proposed that the amount of transfer depends on how many shared elements there are between the learned tasks and the transfer tasks, which is now widely known as the theory of Identical Elements [25]. This theory was later refined by Gick and Holyoak, when they brought out the concept of perceived similarity [26]. Perceived similarity depends on not only the objective number of shared elements, but also the knowledge or expertise of the person performing the transfer task. Essentially, it was pointed out that the more a subject can relate the current transfer tasks to the past learned tasks, and perceive them similarly, the more transfer will take place [26].

In our pyramid experiment, though every subject was presented with the same set of problems, these problems might be perceived very differently. It is likely that the consistency at the neural level in our correlation analysis reflects how similarly the set of new problems appear to the subjects, compared with the trained problems. The more one finds the new and modified problems similar to the trained ones, the more one is able to use knowledge about having solved the trained problems. Though the perceived similarity is not directly measured between the Exception problems encountered during transfer tasks and the Regular problems trained during learning tasks, it can be estimated with brain consistencies among all problems during the transfer tasks. The fact that the subjects are able to respond to different problems in a similar and consistent way reflects how much general learning has already occurred.

HSMM states	Encode	Plan	Solve	Respond	Overall
Regular-Exception	**0.265**	**0.408**	**0.436**	**0.413**	**0.450**
Exception-Exception	**0.231**	**0.402**	**0.442**	**0.416**	**0.424**
Regular-Regular	0.194	**0.302**	**0.287**	**0.306**	**0.367**

Table 3: Pearson correlations between the subject performance and the within-subject brain consistency for each of the 4 states: Encode, Plan, Solve and Respond. Brain consistency is calculated between the Regular and the Exception problems, among the Exception problems, and among the Regular problems. For example, the 0.265 indicates that there is a significant correlation between subject performance and the within-subject brain consistency between the Regular and the Exception problems for the encode state. An overall correlation is also obtained through collapsing the 4 states to a single vector of 56 (14x4) regions. Significant ($p < 0.05$, two-tailed) correlations are in bold.

The next step is to verify if there is a significant correlation between the overall performance and the brain consistency measured between the Regular problems and the Exception problems, and if this effect can be estimated with the brain consistency measured among the Exception problems when there are not enough Regular problems. One fifth of the problems during the scanning session were familiar Regular problems while the rest were new Exception problems. If the brain consistency among all problems does reflect the consistency

between new problems and familiar problems, the same effect should arise if we carry out the correlation analysis between the Exception problems and the Regular problems. Since the number of Regular problems is too small to be analyzed on only the first 10 correctly solved problems, we take into consideration all the problems in the experiments for each subject.

Table 3 shows that there is a significant correlation between subject performance and the within-subject brain consistency between the Regular and the Exception problems for the encode state. It is obtained in a similar way as Table 1 except that the analysis is carried out over the entire experiment on a group of problems of interest. To statistically verify if a linear regression model using 8 predictors from brain consistency using both Regular-Exception pairs and Exception-Exception pairs will produce significantly better prediction than a linear regression model with only 4 predictors from Exception-Exception pairs, an F-test on their Root Sum Squared (RSS) is applied which gives $p = 0.394(> 0.05)$. Meantime, if we compare a model using 8 predictors including both Regular-Exception pairs and Regular-Regular pairs with that using only 4 predictors from Regular-Regular pairs, the F-test gives $p = 0.0137(< 0.05)$. Though in Table 3, using Regular-Exception pairs correlates better to the subject performance than both Exception-Exception pairs and Regular-Regular pairs, adding the Regular-Exception pairs will improve the prediction using only Regular-Regular pairs, but will not improve significantly the prediction using Exception-Exception pairs. This result is in accordance with our hypothesis that the amount of transfer in our task depends on perceived similarity between new Exception problems and familiar Regular problems, and that this effect can be estimated when using only pairs among Exception problems, thereby justifying using brain consistency among all types of problems during the first 10 correctly solved problems.

5.1.5 Observation 5: Within-subject consistency decreases towards the end of the experiment

Another interesting observation from our results is the practice effect. As we examine closer at the magnitude of the within-subject brain consistency across all subjects in Figure 2, we can observe that there is a drop of brain consistency at the late stage of the problem solving compared with the early stage, for both the group of better subjects and the group of weaker subjects. This decreased consistency may account for the decreased correlation between the brain consistency and the overall performance. However, if consistency indicates expertise, with more training, why there would be decrease in its magnitude? It turns out that the perceived similarity between different situations is not the sole factor that gives consistency.

There has been evidence in previous experiments, that behaviors that are more automatic and impulsive are more consistent than behaviors that are relatively controlled and cognitively mediated [27]. In our experiment, the early stage of the experiment when subjects are first exposed to the new problems, their response could be more automatic when they adapt a general strategy developed when solving the Regular problems before the fMRI scanning. As the time goes on, after being exposed to sufficiently many novel Exception problems, the subjects need to develop more diverse and problem-specific strategies as the problem solving process becomes more cognitively controlled. The better subjects not only respond consistently at the early stage of the experiment, but also have a larger decrease in consistency as they develop more problem-specific strategies as is observed in Figure 2. Other than the need to develop refined strategies after more practice, the decreased consistency might also result from its weakened dependence on the perceived similarity.

It has been demonstrated in a learning task that, as time goes on, the proportion of reminding problems where consistency occurs is decreasing [7]. Reminding here is the process where the subjects are able to relate the current task to the learned task. In the same experiment, a similar observation is obtained that is very comparable to our results for performance prediction. The extent that the subjects can relate the current task to the learned one has predictable effect on the task performance. However, this reliance decreases after more practice [7]. A few possible explanations were presented in the paper, such as the increased interference after exposure to more episodes and more goal-oriented retrieval with gained expertise.

5.2 Result 2: The Selected Brain Voxels

Experiment 2 was designed to test whether the better subjects have a higher brain consistency, measured over selected brain voxels, than the weaker subjects.

Illustrated in Figure 5, selected brain voxels during the first 10 correctly solved problems overlapped with some of the 14 pre-defined regions like the angular gyrus, prefrontal cortex, anterior cingulate cortex and Brodmann area 10. The exact number of the selected voxels differs at each run, but is around one tenth of the total number of voxels. As is observed in Figure 6, there is a considerable match between the predicted performance using LOOCV and the actual accuracy of the subjects, with the correlation of the two being 0.640 and MSE being 0.029. LOOCV with selected brain voxels further improves the performance prediction compared with previously using only the 14 ROIs with correlation of 0.561 and MSE of 0.0399. Similar analysis is also done for the LOOCV prediction using only late consistency which gives a correlation to performance of 0.278 and MSE of 0.074.

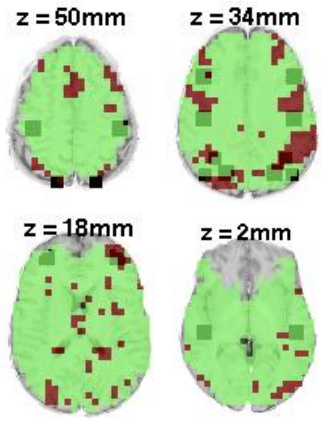

Figure 5: Visualization of the brain voxels in red whose averaged activations over the first 10 correctly solved problems are significantly ($p \leq 0.05$) correlated with the subject overall performance, for a representative state - Encode. The rest of brain voxels are colored light green.

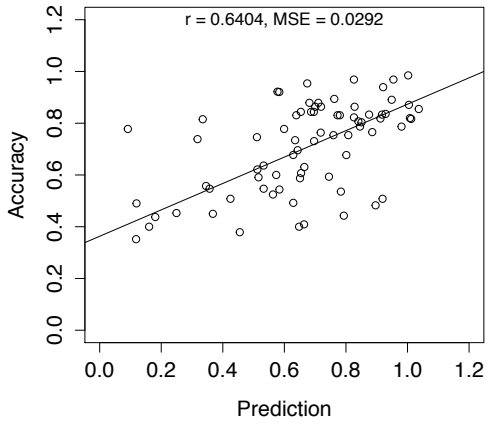

Figure 6: Leave-one-subject-out cross validation (LOOCV) performance prediction of 75 subjects compared with actual subject performance, using the weights of multiple regression analysis obtained from the rest of the subjects. A total of 4 predictors are 4 brain consistencies measured across selected brain voxels among the first 10 correctly solve problems for each of the 4 states. (r = .6404, MSE = .0292)

5.3 Result 3: Spectral Brain Parcellation

Experiment 3 was designed to test whether the better subjects have a higher brain consistency, measured over the entire brain volume, than the weaker subjects.

In this section, a different measure of brain consistency at the level of the whole brain is proposed without a pre-selection procedure on voxels or regions. Though it is not clear whether extending the analysis from selected regions to the whole-brain level would include previously ignored important information or unnecessarily introduce redundant information, it is of interest to verify if the same result still holds from an alternative perspective. In using spectral clustering, brain consistency is measured as the consistency among selected/grouped regions after considering local correlated regions. A representative clustering is shown in Figure 7. Brain activations have been averaged within each clustered regions for better visualization of the cluster boundaries.

	First 10	Last 10
14 ROIs	**0.558**	**0.365**
Clustering(ARI/AMI)	**0.414/0.430**	**0.267/0.255**

Table 4: Averaged pairwise consistency of the first/last 10 trials in correlating with participant performance under two different measures of consistency. Significant ($p < 0.05$, two-tailed) correlations are in bold. It shows that there is a significant correlation between within-subject consistency, measured either over the 14 pre-selected regions or the entire brain, and subject performance.

It can be observed (see Table 4) that there is a significant correlation between within-subject consistency, measured either over the 14 pre-selected regions or the entire brain, and subject performance. To have the results from two different measures of consistency comparable to each other, the Overall correlation between the brain consistency and the

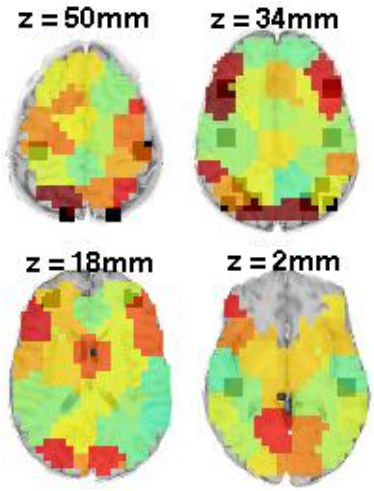

Figure 7: A represented spectral clustsering on the averaged brain activity of subject 1. Brain activations have been averaged within each cluster.

subject performance across the 14 ROIs is used that has already been obtained in Table 1. In both cases, the correlation between the brain consistency and the subject performance is higher for the first 10 correctly solved problems than for the last 10 correctly solved problems. As we found for the measure of within-subject brain consistency for previous sessions, there is a significant difference between this measure brain consistency in better subjects and weaker subjects in one-way between-subjects ANOVA ($p < 0.0049$ for both ARI and AMI in F-test), and significant difference between early consistency and late consistency in one-way within-subjects ANOVA ($p = 0.001$ for RI and $p = 0.0016$ for MI in F-test). This observation is reassuring in two aspects. For one, a whole-brain level analysis ensures that the obtained prediction effect is not dependent on focusing on selected brain regions. For the other, an alternative measure of brain consistency based on brain parcellations ensures that the obtained effect is not dependent on one single measure of consistency such as Pearson correlation.

6 Conclusion

This study has shown that success in extending a human problem-solving procedure from familiar to unfamiliar problems is reflected in how consistent subjects' brain responses are. This consistency refers to both how brain responses of one subject deviate away from the global average (between-subjects consistency), and how consistently subjects respond to different problems (within-subject consistency), with the latter one correlating stronger to the subject overall performance and specific to the transfer task. Within-subject brain consistency is most correlated with the subject performance when examining the early stage of the problem solving, which can serve as an effective neural predictor. During the later stage of the problem solving, subjects start developing problem-specific strategies that decrease the brain consistency over time.

Though previous studies have explored the relation between consistency and subject

performance during transfer of learning, our study is the first in identifying such effect at the level of neural activity and further applying it in an attempt to predict performance. Section 5.2 showed that the accuracy of predicting subject performance has been further improved by selecting the most involved brain voxels than only using the predefined 14 regions. In Section 5.3 we used a data-driven approach to characterize brain consistency at the level of the whole brain, and we found a similar correlation between the brain consistency and the overall subject performance. Thus, all three approaches converge to the same conclusion that subjects who have more consistent brain activation perform better.

We suggest that the relationship between subject performance and brain consistency is due to the perceived similarity between the familiar problems that one has been trained on (Regular problems) and the new problems (Exception problems). The more one finds the new and modified problems similar to the trained ones, the more one is able to use knowledge from having solved the trained problems. This explanation is supported by the observation that there is significant correlation between the overall subject performance and the brain consistency calculated over the entire experiment between the Regular problems and the Exception problems.

References

[1] J. Desse. *Transfer of Training: The Psychology of Learning*. McGraw-Hill, 1958.

[2] R. E. Ripple and D. J. Drinkwater. *Encyclopedia of Educational Research*. Academic Press, 1982.

[3] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

[4] John R. Anderson, Shawn Betts, Jennifer L. Ferris, and Jon M. Fincham. Neural imaging to track mental states while using an intelligent tutoring system. *Proceedings of the National Academy of Sciences*, 107(15):7018–7023, 2010.

[5] Anja Ischebeck, Laura Zamarian, Michael Schocke, and Margarete Delazer. Flexible transfer of knowledge in mental arithmetic — an fmri study. *NeuroImage*, 44(3):1103 – 1112, 2009.

[6] John R. Anderson and Jon M. Fincham. Extending problem-solving procedures through reflection. *Cognitive Psychology*, 74(0):1 – 34, 2014.

[7] Brian H. Ross. Remindings and their effects in learning a cognitive skill. *Cognitive Psychology*, 16(3):371 – 416, 1984.

[8] John R. Anderson and Jon M. Fincham. Discovering the sequential structure of thought.

[9] Gary H. Glover. Deconvolution of impulse response in event-related {BOLD} fmri1. *NeuroImage*, 9(4):416 – 429, 1999.

[10] K. J. Friston. *Statistical parametric mapping : the analysis of funtional brain images*. Elsevier/Academic Press, Amsterdam; Boston, 2007.

[11] B Obermaier, C Guger, C Neuper, and G Pfurtscheller. Hidden markov models for online classification of single trial {EEG} data. *Pattern Recognition Letters*, 22(12):1299 – 1309, 2001. Selected Papers from the 11th Portuguese Conference on Pattern Recognition - {RECPAD2000}.

[12] H. Wainer and S. Messick. *Principles of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*. Taylor-Francis, 1983.

[13] Shun zheng Yu and Hisashi Kobayashi. Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden markov model. *Signal Processing, IEEE Transactions on*, 54(5):1947–1951, May 2006.

[14] Shun zheng Yu and Hisashi Kobayashi. An efficient forward-backward algorithm for an explicit-duration hidden markov model. *Signal Processing Letters, IEEE*, 10(1):11–14, Jan 2003.

[15] J. Borst and J. Anderson. The discovery of processing stages: Analyzing eeg data with hidden semi-markov models. *NeuroImage*, 2014.

[16] Henning Gibbons and Jutta Stahl. Response-time corrected averaging of event-related potentials. *Clinical Neurophysiology*, 118(1):197 – 208, 2007.

[17] Jose Luis Perez Velazquez, Richard Wennberg, Matthias Ihrke, Hecke Schrobsdorff, and J. Michael Herrmann. *Springer Series in Computational Neuroscience*, volume 2, pages 165–189. Springer New York, 2009.

[18] Edna C. Cieslik, Karl Zilles, Svenja Caspers, Christian Roski, Tanja S. Kellermann, Oliver Jakobs, Robert Langner, Angela R. Laird, Peter T. Fox, and Simon B. Eickhoff. Is there "one" dlpfc in cognitive action control? evidence for heterogeneity from co-activation-based parcellation. *Cerebral Cortex*, 23(11):2677–2689, 2013.

[19] Jason W. Bohland, Hemant Bokil, Cara B. Allen, and Partha P. Mitra. The brain atlas concordance problem: Quantitative comparison of anatomical parcellations. *PLoS ONE*, 4(9):e7200, 09 2009.

[20] Torgny Greitz, Christian Bohm, Sven Holte, and Lars Eriksson. A computerized brain atlas: Construction, anatomical content, and some applications. *Journal of Computer Assisted Tomography*, 15(1), 1991.

[21] Andreia V. Faria, Suresh E. Joel, Yajing Zhang, Kenichi Oishi, Peter C.M. van Zjil, Michael I. Miller, James J. Pekar, and Susumu Mori. Atlas-based analysis of resting-state functional connectivity: Evaluation for reproducibility and multi-modal anatomyâfunction correlation studies. *NeuroImage*, 61(3):613 – 621, 2012.

[22] Sean L Simpson, Robert G Lyday, Satoru Hayasaka, Anthony P Marsh, and Paul J Laurienti. A permutation testing framework to compare groups of brain networks. *Frontiers in Computational Neuroscience*, 7(171), 2013.

[23] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, 11:2837–2854, December 2010.

[24] J. Anderson, D. Bothell, J. Fincham, and J. Moon. The sequential structure of brain activation predicts skill. *Journal of Cognitive Neuroscience*, 2014.

[25] E.L. Thorndike and R.S. Woodworth. The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review*, 8:247–261, 1901.

[26] M. L. Gick and K.J. Holyoak. *The Cognitive Basis of Knowledge Transfer*. Academic Press, 1987.

[27] R.Michael Furr and David C. Funder. Situational similarity and behavioral consistency: Subjective, objective, variable-centered, and person-centered approaches. *Journal of Research in Personality*, 38(5):421 – 447, 2004.

[28] Aviv Mezer, Yossi Yovel, Ofer Pasternak, Tali Gorfine, and Yaniv Assaf. Cluster analysis of resting-state fmri time series. *NeuroImage*, 45(4):1117 – 1125, 2009.

[29] Shing-Chung Ngan and Xiaoping Hu. Analysis of functional magnetic resonance imaging data using self-organizing mapping with spatial connectivity. *Magnetic Resonance in Medicine*, 41(5):939–946, 1999.

[30] Cyril Goutte, Peter Toft, Egill Rostrup, Finn Ã. Nielsen, and Lars Kai Hansen. On clustering fmri time series. *NeuroImage*, 9(3):298 – 310, 1999.

[31] Bertrand Thirion, Gael Varoquaux, Elvis Dohmatob, and Jean-Baptiste Poline. Which fmri clustering gives good brain parcellations? *Frontiers in Neuroscience*, 8(167), 2014.

[32] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, Aug 2000.

[33] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(11):1101–1113, Nov 1993.

[34] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[35] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[36] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[37] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1073–1080, New York, NY, USA, 2009. ACM.

7 Appendix

7.1 Spectral clustering

Brain parcellation techniques are used to group the brain voxels into a set of non-overlapping regions. It accounts for the functionally homogeneity of brain voxels from the data instead of implementing any anatomical priors. Methods in the literature includes spectral clustering, K-means [28], self-organizing maps [29], and hierarchical clustering [30]. Notably, spectral clustering has been very successful in a couple of recent studies due to its capability to incorporate spatial constraints that will lead to spatially connected components of reasonable and balanced sizes [31]. Before its popularity in brain imaging analysis, it was earlier demonstrated in an image segmentation task to be able to effectively extract the global impression instead of low-level cues of an image [32], which is in accordance of our purpose of characterizing brain response on the level of whole brain instead of individual voxels. With spectral clustering, brain parcellation is essentially re-formulated into a graph-partition problem. In particular, fMRI is now represented as an undirected weighted similarity graph $G = \big(V, E\big)$, where the set of nodes V includes all the brain voxels, and the set of edges E includes edges between pairs of neighbor voxels. Weights of the edges are defined in a weighted adjacent matrix W as below:

$$w_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2/(2\sigma^2)), \text{ if } \|x_i - x_j\| \leq \epsilon \\ 0, \text{ otherwise} \end{cases}$$

To incorporate the brain responses across different states, x_i refers to a vector of brain activations of the i^{th} voxel that corresponds to 4 mental states. ϵ defines what are considered adjacent locations in the graph G.

In a simplified case where the graph is to be partitioned into two disjoint sets A and B, the disassociation between the two sets can be summarized in the total weight of removed edges, which is called the *cut*:

$$cut(A, B) = \sum_{u \in A, v \in B} w(u, v)$$

A successfully partitioning would be to minimize the *cut* value so as to maximize the inter-set difference. With the same criteria, the problem of k-subgraph partitioning can be solved by recursively finding the minimum cut that bisects the existing segments in a clustering method proposed by Wu and Leahy [33]. However, the minimum cut criteria increases with the number of edges across two sets of interest, thus favoring small isolated sets sometimes as a single node. To avoid such partitioning, a modified criteria is used which is referred to as the *normalized cut* (Ncut):

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)}$$

where $assoc(A, V) = \sum_{u \in A, x \in V} w(u, x)$ represents the total connection between nodes in A and all the rest of nodes in the graph G. It is not difficult to notice that:

$$cut(A, B) = assoc(A, V) - assoc(A, A) = assoc(B, V) - assoc(B, B)$$

Minimizing this Ncut would be equivalent to maximizing the disassociation between the two sets. Small isolated sets S would no longer have small Ncut since the total connection $assoc(S, V)$ will be small too. Interestingly, from above, we can re-write Ncut as below:

$$Ncut(A, B) = 2 - (\frac{assoc(A, A)}{assoc(A, V)} + \frac{assoc(B, B)}{assoc(B, V)}) = 2 - Nassoc(A, B)$$

where $Nassoc(A, B)$ is intuitively the sum of association within groups. In other words, when minimizing the inter-set disassociation, intra-group association is also maximized. This is a second advantage introduced by Ncut other than avoiding small isolated sets [32].

The optimized partition could be obtained by solving the generalized eigenvalue system:

$$(D - W)v = \lambda Dv$$

Spectral clustering solving an eigenvalue system corresponds to minimizing the *cut*, whereas a normalized spectral clustering solving the generalized eigenvalue system corresponds to minimizing the $Ncut$. The above distinction of normalized and unnormalized cut justifies our use of spectral clustering with generalized eigenvalues where we will obtain reasonable brain partitions with comparable cluster sizes. The detailed steps of the normalized spectral clustering is summarized as below:

Algorithm

- Similarity graph G is constructed with W being its weighted adjacent matrix
- Compute the graph Laplacian matrix $L = D - W$
- Compute k generalized eigenvectors v_1, v_2,..., v_k that correspond to the k smallest eigenvalues of $Lv = \lambda Dv$
- Construct a matrix $V \in R^{\text{nxk}}$ with columns being v_1, v_2,..., v_k
- Apply k-means algorithm on rows of V to obtain clusters C_1, C_2,..., C_k

σ in the Gaussian similarity function controls the width of the neighborhood. If sigma is too small, the graph will not be safely connected to apply spectral clustering; whereas if it is large, the Gaussian similarity function will be too flat to capture any similarity. It has been suggested in a previous application of spectral clustering that a practical guide on the choice of σ is around the scale of the average distances between connected voxels which is around 0.01 in our case [31]. The number of clusters, k, need to be specified beforehand. $k = 100$ is used since anatomical atlases usually propose a decomposition into about 100 regions [31]. ϵ is 7 (the size of a brain voxel being 2). This choice of ϵ is large enough to guarantees the local connection, while at the same time small enough to not give rise to a graph too large to have spectral clustering done in a reasonable amount of time.

7.2 Measures for Comparing Clusterings

To extend the use of brain parcellation to characterize the changes of brain responses across distinct experimental conditions, measures for comparing clusterings can be applied to indicate the similarity/consistency of two brain responses. Two popular classes of measures are pair-counting based and information-theoretic based [23].

For the first class, the most used measure is Rand Index (RI), which characterizes the consistency of two clustering, U and V, by examining the co-occurrence of all the possible pairs. In specific, let n_{00} represents the number of pairs that are present in different clusters in both U and V; let n_{01} represent the number of pairs that are in different clusters in U but the same cluster in V; let n_{10} represent the number of pairs that are in the same cluster in U but in different clusters in V; let n_{11} represent the number of pairs that are in the same cluster in both U and V. Rand Index is thus defined intuitively as the proportion of pairs that consistently show up together or separately in two ways of clustering [34]:

$$\text{RI(U,V)} = \frac{n_{00} + n_{11}}{\binom{N}{2}}$$

Despite that RI(U,V) has a range from 0 to 1, practically the range is much narrower because Rand Index of two clusterings by chance is larger than 0. To facilitates interpretation and comparison across different conditions, correction for chance is necessary. It is suggested that the general form of a similarity index corrected for chance is [35]:

$$\text{Adjusted Index} = \frac{\text{Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}}$$

which gives rise to the use of Adjusted Rand Index(ARI) [35]:

$$\text{ARI(U,V)} = \frac{2(n_{00}n_{11} - n_{01}n_{10})}{(n_{00} + n_{01})(n_{01} + n_{11}) + (n_{00} + n_{10})(n_{10} + n_{11})}$$

The second class builds upon fundamentals in information theory [36]. Given two clusterings U and V, mutual information (MI) measures how much knowing one of the clustering can inform us about the other, which can be used as an indicator of how similar/consistent two clusterings are. It could be quantified by how much uncertainty of U is reduced upon observing V, which can be expressed as below:

$$\text{MI(U,V)} = H(U) - H(U|V) = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{n_{ij}}{N} log \frac{n_{ij}/N}{a_i b_j / N^2}$$

after using the facts that

$$H(U) = -\sum_{i=1}^{R}\frac{a_i}{N} log \frac{a_i}{N}$$

$$H(U|V) = -\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{n_{ij}}{N} log \frac{n_{ij}/N}{b_j/N}.$$

$A = \{a_i\}$ where i = 1,2,3... $B = \{b_j\}$ where j = 1,2,3 ... are two clusterings on the same set of data. n_{ij} represents the overlapping elements between cluster a_i in clustering A and

cluster b_j in clustering B. Similarly, to adjust MI for chance [37], we have the Adjusted Mutual Information(AMI):

$$\text{AMI(U,V)} = \frac{I(U,V) - E\{I(U,V)\}}{\max\{H(U), H(V)\} - E\{I(U,V)\}}$$

So that by chance AMI would be 0 and the largest one can obtain is 1.

7.3 14 Pre-defined Regions

	ROIs
1	Fusi
2	Aural
3	Caud
4	PFC
5	Vocal
6	Parietal
7	ACC
8	Mot
9	AntIn
10	MidIn
11	BA10
12	AngGyr
13	Hips
14	PSPL