Title
From "where" to "how": computational approaches to assessing information representation in neuroimaging studies of higher cognition

Author
Chiang, Jeffrey

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

From "where" to "how": computational approaches to assessing information representation in neuroimaging studies of higher cognition

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Psychology

by

Jeffrey Chiang

2018

ABSTRACT OF THE DISSERTATION

From "where" to "how":

computational approaches to assessing information representation

in neuroimaging studies of higher cognition

by

Jeffrey Chiang

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2018

Professor Martin M. Monti, Chair

While a number of sophisticated computational and theoretical models exist for human behavior in the cognitive science literature, the relationship between these models and the underlying neural computation has rarely been explored. The aim of this thesis is to propose novel applications of neuroimaging analysis methods combined with explicit modeling to bridge the gap between computational models and cognitive neuroscience, specifically in studies of higher cognition. In Chapter 1, I provide the necessary methodological background to these projects, describing in detail current univariate and multivariate approaches to functional MRI (fMRI) analysis. I then describe three approaches, *cross-classification*, *representational similarity analysis*, and *encoding analysis* that allow claims about the underlying representations and computations to be made in neuroimaging studies. In Chapters 2, 3, and 4, I present experiments using these approaches, showing how they allow us to arbitrate between different theories of the representations and computations underlying higher

cognition, building upon prior localization work. Finally, in Chapter 5 I propose a new computational framework for encoding analyses that allows for directly integrating computational models with neuroimaging analysis.

The dissertation of Jeffrey Chiang is approved.

Ying Nian Wu

Keith Holyoak

Hongjing Lu

Martin M. Monti, Committee Chair

University of California, Los Angeles

2018

To my grandparents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I've had the pleasure of being surrounded by some remarkable people during my time at UCLA.

First and foremost, I thank those on the front lines: my parents for always being ready to support and listen, and my brother Jonathan for understanding my constant complaints about computing power. And to my partner Kyra, your patience and understanding as I've worked through my program is far beyond what I deserve. From critiquing my poster designs to sending the perfectly appropriate web comic whenever I've been struggling, I truly appreciate everything you've done to help me make it through this program.

I am also incredibly lucky and humbled to have trained under some of my greatest role models. Dr. Martin Monti, I'm not sure even where to begin. You have given me first the tools and then the freedom to pursue my academic interests, all the while leading by example as an incredible researcher and mentor. Despite your productivity, you've shown a kindness and willingness to always be there that is not lost upon me, and I will always be proud to boast about how great of an advisor I've had. Dr. Hongjing Lu, you helped lay the foundations for my research career when I was just an undergraduate research assistant in your lab, and it has been an honor to learn from and eventually work alongside you. Dr. Keith Holyoak, I have always been in awe of your wealth of knowledge and insights. On top of that, it has been a pleasure to learn from your perspective and wit over the last few years, and every conversation with you has inspired me to further broaden my knowledge. Dr. Ying Nian Wu, much of the later work in this dissertation would not have been possible without

the statistical intuition you've provided me. You've challenged me to approach everything with methodological rigor and it shows whenever I discuss recent developments with my colleagues.

If I've learned anything during the last few years, it is that being surrounded by good people is the key to good work. I'd especially like to thank Dr. Evan Lutkenhoff, for always having the answers to my incredibly obscure methodological questions as well as general advice for surviving graduate school. I'd also thank Dr. Julia Crone, Micah Johnson, John Dell'Italia, and the rest of the Monti Lab for all of their contributions and support.

Many of the ideas in these projects were born from conversations with Drs. Nicco Reggente, Alex Huth, Majed Samad, James Kubricht, and Uri Maoz. I am grateful that they were always willing to lend a critical ear.

Finally, thank you to all of my collaborators (below). It's been a pleasure to learn about your respective subfields through our projects and I hope my future work is even half as enjoyable as my time working with you.

Chapter 2 is a version of: Chiang, J.N., Rosenberg, M., Bufford, C., Stephen, D., Lysy, A., Monti, M.M. The language of music: Common neural codes for structured sequences in music and natural language. *Brain and Language*, 2018. MMM and AL developed the study concept and secured the funding. MMM, MHR, CAB, and DS devised the study design. MHR conducted behavioral testing and neuroimaging data collection. JNC and MHR performed the data analysis and, together with MMM, interpreted the results. MMM drafted the manuscript, MHR and JNC provided critical revisions. All authors contributed to subsequent editing of the manuscript. This research was supported by the UCLA OVCR-

COR Transdisciplinary Seed Grant "Language, Music, and the Brain" to AL and MMM

Chapter 3 is a version of: DeWolf M., Chiang J.N., Bassok M., Holyoak K.J., Monti M.M. Neural Representations of Magnitude for Natural and Rational Numbers. *Neuroimage*, 141, 304-312. 2016. MD, KJH, MMM developed the study concept and design. JNC and MD collected and analyzed the data. MD wrote the manuscript, with theoretical contributions from MB and KJH and techinical contributions from JNC and MMM.

Chapter 4 is a version of: Chiang, J.N., Peng, Y., Lu, H., Holyoak, K.J., Monti, M.M. Neurocomputational models predict brain activity during analogical reasoning. (in preparation). JNC collected and analyzed the data, with additional analyses from HL and YP. JNC and KJH and drafted the manuscript.

Chapter 5 is a version of: Chiang, J.N., Monti, M.M. Incorporating priors from computational models in neuroimaging encoding methods: Tikhonov regression. *bioRxiv*. 2018. JNC performed the analysis and drafted the manuscript, with revisions from MMM. It is also in preparation for journal submission.

The work described in this dissertation was supported by the Staglin IMHRO Center for Cognitive Neuroscience at UCLA, the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program, the Neuroimaging Training Program (NITP: NIH-RFA DA-06-011), and the UCLA Graduate Division via the Distinguished University Fellowship and Graduate Summer Research Mentorship.

VITA

2011	Undergraduate Student Research Program (USRP) Fellow, NASA-Jet Propulsion Laboratory
2012	B.S. (Cognitive Science), minor (Statistics), Computing specialization, UCLA
2013-2014	Distinguished University Fellowship, UCLA Department of Psychology
2014	Graduate Summer Research Mentorship Award, UCLA Graduate Division
2014	M.A. (Psychology), UCLA
2014-2015	NeuroImaging Training Program Fellow (NITP- NIH-RFA DA-06-011)
2014-2017	Teaching Assistant, Department of Psychology, UCLA
2015-2018	National Defense Science and Engineering Graduate (NDSEG) Fellow, Department of Defense
2017-2018	Teaching Assistant, Anderson School of Management, UCLA
2018	Teaching Consultant, Chapman University

PUBLICATIONS AND PRESENTATIONS

* denotes joint first authors

Chiang, J.N.*, Rosenberg, M.*, Bufford, C., Stephen, D., Lysy, A., Monti, M.M. The language of music: Common neural codes for structured sequences in music and natural language. *Brain and Language*, 2018.

DeWolf M., **Chiang J.N.**, Bassok M., Holyoak K.J., Monti M.M. Neural Representations of Magnitude for Natural and Rational Numbers. *Neuroimage*, 141, 304-312. 2016.

Thurman S.M., van Boxtel J.J., Monti M.M., **Chiang J.N.**, Lu H. Neural adaptation in pSTS correlates with perceptual aftereffects to biological motion and with autistic traits. *Neuroimage*, 136:149–161, 2016.

Morales, J, **Chiang, J.N.** and Lau, H. Controlling for performance capacity confounds in neuroimaging studies of conscious awareness. *Neuroscience of Consciousness* (2015).

Lutkenhoff, E., **Chiang, J.**, Tshibanda, L., Kamau, E., Kirsch, M., Pickard, J., Laureys, S., Owen, A., Monti, M. Thalamic and extrathalamic mechanisms of (un)consciousness after severe brain injury. *Annals of Neurology* (2015).

Lutkenhoff E.S., Rosenberg M., **Chiang, J.**, Zhang R., Pickard J.D., Owen A.M. and Monti M.M. Optimized brain extraction script for pathological brains (optiBET). *PLoS One* (2014).

Monti, M.M., **Chiang, J.**, Lutkenhoff, E.S. "Thalamic Atrophy and Awareness in Disorders of Consciousness." *Society for Neuroscience* (2013).

Chiang, J., Zhang, Y., Lu, T., Chao, T-H., Composite wavelet filters for enhanced automated target recognition. *SPIE Proceedings*, Vol 8398, 14, 2012.

INTRODUCTION

The introduction of machine learning techniques to neuroimaging analysis (Haxby et al., 2001) has led to new and profound understandings of the brain. Early studies used multivariate pattern (MVP) analyses to expand upon neuropsychological and animal model findings with specific hypotheses. The early promise of MVP analyses was that by separating cognitive or mental states via classification, they could properly address "how", and not "where" information was represented in the brain.

However, as these methods became more widely adopted, it has become apparent that on their own, MVP methods do not fulfill this promise. In contrast to the early, hypothesis driven classification studies, recent studies have taken an "exploratory" flavor, applying big data approaches (e.g., data mining and machine learning methods) to probe neural representations (see Cohen et al., 2017; Hebart & Baker, 2017 for specific examples). On top of separating cognitive states using classification, many studies interpret classifier performance post-hoc (see reverse inference, Poldrack, 2006). From a neurocognitive standpoint, it has been shown that these methods still only address "where" information is represented, except from an information-theoretic framework (Haynes, 2015; Hebart & Baker, 2017), bringing us no closer to understanding the underlying computation using these data-driven methods.

On the other hand, a number of sophisticated computational and theoretical models have been proposed in the cognitive psychological literature. These models do provide computational or mechanistic explanations of behavior, but the relationship between these models and the underlying neural computation has rarely been explored in fields other than visual

perception (Agrawal, Stansbury, Malik, & Gallant, 2014).

The aim of this thesis is to demonstrate novel applications of MVP methods combined with explicit modeling to bridge the gap between the theoretical models from psychology and cognitive neuroscience, specifically in studies of higher cognition. In Chapter 1, I provide the necessary methodological background to these projects, describing in detail the multivariate and encoding analyses that allow for new insights on the neural computations. In Chapter 2, I use multivariate pattern classification (MVPC) in a *cross classification* approach in order to study shared neural representations across language and music. In Chapter 3, I apply *representational similarity analysis* (RSA- Kriegeskorte, Mur, & Bandettini, 2008) to arbitrate between theories of number representation in the parietal cortex. Chapter 4 uses a combination of classification, RSA, and encoding analyses to propose a neurocomputational account of analogical reasoning. Finally, in Chapter 5 I propose an extension to a popular analysis framework (Naselaris, Kay, Nishimoto, & Gallant, 2011) that allows for directly integrating computational models into neuroimaging analysis.

CHAPTER 1

Task based fMRI analysis

Functional Magnetic Resonance Imaging (fMRI) aims to measure brain metabolic response (Brain Oxygen Level Dependent- BOLD) across time (Ogawa, Lee, Kay, & Tank, 1990). In a typical (1.5, 3, or 7 Tesla) MRI machine, subjects undergo a scanning session, typically an hour long broken into several separate scans, which will be henceforth referred to as "runs". When coupled with an experimental task, the fMRI activity during different experimental conditions can be compared to make inferences about brain activity.

Preprocessing Given the nature of the scanning equipment, human tendencies and hemodynamic response (as understood by Buxton, Wong, & Frank, 1998; Boynton, Engel, Glover, & Heeger, 1996, there are a number of standard processing techniques that are shared across fMRI studies of healthy subjects in order to maximize signal to noise. As MR images are highly susceptible to motion artifacts (which manifest as signal changes orders of magnitude higher than signal changes associated with behavior, Jenkinson, Bannister, Brady, & Smith, 2002), raw data are typically motion corrected and temporally filtered. In order to make cross-subject generalizations, data from each individual subject are aligned into a standard or average template. In the univariate framework, spatial smoothing is commonly applied to satisfy the assumptions of the model and to account for small heterogeneities in anatomy

from subject to subject (Monti, 2011). However, this technique is decreasingly common in multivariate designs (Cohen et al., 2017), as these local heterogeneities can be very informative in a multivariate context. Finally, the BOLD signal is well known to be a slow response profile. It is common to model this lag using a hemodynamic response function (HRF) either using assuming a specific shape on the function, or non-parametrically modeling this lag using multiple basis functions (e.g. Stephan, Harrison, Penny, & Friston, 2004; Mumford, Turner, Ashby, & Poldrack, 2012).

1.1 Univariate analyses

The traditional "massive univariate" approach to fMRI time-series analysis assumes some hemodynamic lag structure, and convolves this with the experimental design in order to fit a general linear model (Haynes, 2015; Monti, 2011) for a summary and critique). In these analyses we typically employ a condition-coded design, in which regressors serve as indicators as to which experimental condition the subject was exposed to at a given timepoint. The resulting condition weight estimates (i.e. β's) are essentially HRF-weighted condition means, and are passed to T-tests with contrasts specified by a contrast matrix (Posner & Deheane, 1994). The contrasts of interest rely heavily on the assumption of cognitive subtraction, relying on experimental design to isolate the cognitive function of interest via subtraction. This was and still is a prevailing strategy in fMRI analysis, capitalizing on the subtraction assumption and clean experimental design. (Monti, 2011). These analyses apply "traditional" statistical inference (i.e. evaluating t-statistics on activation differences between condition averages), employing multiple comparison correction for valid inference. A crucial point in

the univariate analysis is that the GLM model is fit over the entire data set, as inference is being conducted on the magnitude of activation parameters (as opposed to model fit, which will be discussed later). In experiments in which multiple levels of conditions are shown, the parametric univariate design (Braver et al., 1997) matrix (with values indicating the "level" of each condition) can be employed.

After group aggregation (described above), voxels exhibiting a contrast significantly greater than zero are termed "activations". Within this framework, a significant "activation" indicates that a cognitive process of interest is associated with increased (or decreased) activity within a specific region.

1.2 Multivariate analyses

Multivariate pattern analyses (sometimes multi-voxel pattern or MVP analyses) encompass a broad range of techniques that assume that patterns of voxel activity, rather than the activity of a single voxel across time, encode meaningful information (Kriegeskorte et al., 2008; Haxby, Connolly, & Guntupalli, 2014; Haynes, 2015; Hebart & Baker, 2017; Cohen et al., 2017). There are two primary approaches to MVP analyses that reflect different information philosophies- one primarily driven by machine learning and by representational geometry.

Additional preprocessing: Parameter estimation MVP and encoding analyses aim to characterize trial-by-trial variability rather than condition-by-condition differences, nullifying the signal to noise benefit of the condition coded univariate analysis. Typically, experimental

trials span several timepoints. While using the time-series of each trial is a viable method-(see (Turner, Mumford, Poldrack, & Ashby, 2012)), in these frameworks it is more common to summarize each trial as a single parameter estimate representing the amplitude or magnitude of activation. This has been accomplished either by averaging the TRs with the highest expected activation given a priori knowledge of the hemodynamic response function (e.g. Mitchell et al., 2008; Haxby et al., 2001; Rissman, Gazzaley, & Mark, 2004) or a GLM approach in which a trial indicator matrix is convolved with an HRF, effectively computing a weighted average of the timepoints comprising each trial (weighted by the shape of the HRF at that time). Different GLM approaches have been explored: the two most common approaches are Least-Squares All (LSA), in which all parameter estimates are simultaneously calculated via GLM, and Least-Squares Single (LSS) in which each parameter estimate is calculated iteratively, with its own GLM regression. Whereas the LSA procedure is proven to provide unbiased estimates of activation amplitude, the LSS procedure trades a slight bias for significant variance reduction (Mumford et al., 2012). Regardless of the procedure, parameter estimates (aka "betas") are concatenated to form a "beta-series" that serves as the input to trial-by-trial analyses (Rissman et al., 2004).

Throughout this thesis, the primary dependent measure is fMRI signal that has undergone this preprocessing, that is, fMRI signal that has been preprocessed and parameterized into condition or trial-by-trial activity. Parameterized data is typically fed into one or more of the analyses below.

1.2.1 MVPC

Multivariate pattern classification (MVPC), at present, is the most commonly used flavor of MVP analysis. If multi-voxel activity patterns can be visualized in a high dimensional space, MVPC aims find some boundary that separates points belonging to different cognitive states or experimental conditions (see Figure 1 from Haxby et al., 2014). This is in contrast with the univariate paradigm described above, in which cognitive states or experimental conditions serve as the predictors for brain response which has implications for implementation and inference (discussed below and in Naselaris et al., 2011, Figure 4). Many strategies for finding this boundary exist (e.g. Alpaydin, 2009, though see Haynes, 2015 for a justification of popular choices), and these general-purpose methods are directly applied to preprocessed fMRI data. The boundary (i.e. classifier) is then evaluated on a test set by predicting the data or labels, and generally a classification accuracy or confusion matrix serves as the dependent measure. A successful classification is taken to mean that the pattern of voxel activity is meaningful to representing the cognitive states of interest. Generally, machine learning techniques can be directly adapted to preprocessed neuroimaging data. However, a primary characteristic of experimentally acquired data is the relatively low number of trials, which gives rise to the problems below.

A common problem in the machine learning literature is rank-deficiency, which generally manifests in MVPC when the number of features (in this case voxels) is greater than the number of samples (in this case trials). Nearly all linear classification methods involve an inversion of the feature-by-feature variance-covariance matrix, which is ill-defined when data are rank-deficient. Unfortunately, in most MVPC problems in which the brain response

serves as the feature set and the trial labels serve as samples, the problem is almost always rank-deficient. The support vector machine (SVM) classifier is typically used due to its robustness to the case in which the number of trials is less than the number of features (Chang & Lin, 2011), but it is not enough to overcome the rank-deficiency problem when running a whole brain classification. Common solutions for this problem are addressed below.

Feature Selection The problem can be remedied by reducing the feature set, i.e. "selecting" voxels that we believe to be implicated and therefore reducing the rank of the feature space. This manifests in several popular approaches: dimension reduction (Alpaydin, 2009), the ROI based approach, in which a priori hypotheses direct the analysis to a small region of the brain (e.g. (Kriegeskorte et al., 2008; Haxby et al., 2001; Friederici, 2011) , and a searchlight approach (Kriegeskorte, Goebel, & Bandettini, 2006) in which many overlapping ROI classifications which span the entire brain volume are run.

Dimension reduction is a general-purpose data driven approach to reducing the number of features by combining redundant information. This technique typically uses linear methods such as Principal Component Analysis (PCA) to rotate the data into a new, more concise basis set that is more efficient at explaining its variability. Another class of dimension reduction techniques involves "screening" variables for classification potential, for example running an F-test to notice differences in conditions (Hanke, Halchenko, Sederberg, Hanson, et al., 2009; Hanke, Halchenko, Sederberg, Olivetti, et al., 2009). However, these methods need to be carefully implemented, as accidentally including validation samples in the dimension reduction algorithm can greatly bias classification results (see *double dipping*).

ROI approaches have already been mentioned, in which prior research or theory motivates

the researcher to isolate their analysis to smaller regions of the brain (e.g. Allen et al 2014). Dimension reduction techniques outlined above can also be combined with an ROI approach to maximize classification performance, with the same double dipping caveats as above.

Searchlight approaches represent a compromise between the hypothesis-driven ROI approach and a whole-brain exploration (Kriegeskorte et al., 2008). This method involves iterating through many small ROIs spanning the entire brain volume, mapping the score (typically classification accuracy) to the center voxel of the ROI. Thus the searchlight approach reveals clusters of local classifiability—a significant cluster in such an analysis would indicate that the surrounding area represents conditions in an information-theoretically meaningful way.

Double Dipping The "crisis of replicability" (e.g.Poldrack et al., 2017) has brought significant attention to MVP analyses, given their seeming ability to find any statistically significant result. It is important to consider "double dipping" (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009), an implementation error in which analysts include either signal from the testing data when fitting the model, or signal from training data when testing the model. This error occurs because the model is "overfit" with respect to the training data and generalizes poorly, and manifests in obvious as well as extremely subtle ways in fMRI analysis.

The most obvious (and most rare) error is to mistake model training error for classification performance. That is, after a model is fit on some dataset, it is run on the same dataset and the result is interpreted. As the model was designed to minimize error on the training dataset, it is no wonder that performance would be at its upper limit when evaluated on the

same set. While this mistake as presented is not common, quantitative "effect" in the training dataset can "bleed into" the testing dataset without careful partitioning. For example, if the entire dataset is standardized prior to model fit (a typical step in MVP analyses), the mean of the testing dataset will be "exposed" to the model when training and artificially boost validation performance while the final model generalizes poorly. This most often happens when researchers try to "balance" the number of trials in each condition in the training and validation sets, unknowingly mixing the two together such that the test set contains data from the same runs as the training trials.

A related problem common to all fMRI analyses is that of temporal auto-correlations. It is obvious through biology and time-series analyses that timepoints are not independent of each other, and care must be taken when splitting training and testing sets such that the same temporal dynamics are not shared between training and testing (Mumford, 2014; Etzel & Braver, 2013). While parameter estimation reduces the likelihood of a trial "bleeding into" the signal of another trial, the truly safe way to split training and testing sets is to leave an entire scanning session (i.e. "run") out for testing, at the cost of losing the power of having that run in the fitted model (Etzel & Braver, 2013).

Limits on inference Given that all data for a single subject are typically acquired at once, the typical "prediction" approach of machine learning is not applicable. That is, we generally cannot train the classifier on one session and predict response in real-time, and cross-subject variability makes between-subject designs difficult in practice. In order to assess whether performance is statistically significant, the best practice is to run a "validation" or "cross-validation" scheme (Etzel & Braver, 2013) in which 1) the data are partitioned into a

"training" and "testing" set, 2) all model fitting (including standardization) is carried out on only the training set 3) the classifier is evaluated on the "testing" set, and 4) this procedure is optionally repeated for many different train/test splits, after which performance is averaged.

Using the cross-validation and validation approaches, we realize limits on the inferences that can be drawn from the classification design devoid of hypotheses. Considering the cost of scanning, time cost of experimental design and the high dimensionality of the data, any decrease in power with any validation procedure strongly impacts inference ability. In fact, different cross-validation procedures have different impacts on variance (Pedregosa et al., 2011).

It is also important to keep in mind that these classifiers are sensitive to any meaningful variation in the data. While one might be able to infer where cognitive states might be represented, MVPC brings us no closer to understanding exactly how the region operates in order to represent such information. A recent perspective has framed MVPC as redundant, rather than complementary to univariate analyses as localization in an information theoretic framework versus activation-based framework (Hebart & Baker, 2017). Crucially, the driver of a successful classification can have a number of indistinguishable explanations (see Figure 2 and Figure 4, left and middle panels from Naselaris & Kay, 2015, making it impossible to truly claim causes and perpetuating the "crisis of replicability" (Poldrack et al., 2017). This ambiguity means that above-chance performance in these studies is not guaranteed to reflect cognition (as it may correspond to other task-related behavior) and is not uniquely related with effect size (Hebart & Baker, 2017). Because the distribution of classifier model weights is unexplored, there is also no principled or parametric approach to conducting

statistical inference on "significant" contributors to classification (Haynes, 2015). As this thesis is concerned with experiments regarding higher cognition, these problems are especially apparent. The psychological dimensions of the stimuli cannot be tightly controlled, and thus unexplained factors can affect variance structure in unknown ways.

Proper use of MVPC With this assessment, I do not wish to discourage the use of MVPC outright. Instead, there is a recent call to use MVPC as the first pass analysis, given its interpretation as an information-theoretic localizer (Haynes, 2015; Hebart & Baker, 2017). Framed this way, MVPC can provide converging evidence with a hypothesized model. In one example, Thurman, van Boxtel, Monti, Chiang, & Lu, 2016 used MVPC to show that local versus global biological motion features could be represented in the posterior superior temporal sulcus (pSTS), and followed up the finding with a hypothesis driven computational investigation on whether subjects' Autism Quotient scores could predict neural adaptation in the region, further suggesting that these differences were the drivers of classification.

Cross classification: Representational inference using MVPC While proper cross-validation and statistical techniques have been developed for MVPC (Etzel & Braver, 2013), the information based inferences of MVPC are not supplemented by experimental design in the same way as univariate analyses. This makes it impossible to determine, in a traditional classification paradigm, the driver of the successful classification. In fact, from an information theoretic standpoint, meaningful representation does not even guarantee high classification accuracy as it does not have a one-to-one relationship with effect size (Hebart & Baker, 2017).

MVPC analyses generally employ cross-validation (Etzel & Braver, 2013; Hebart & Baker, 2017; Cohen et al., 2017), which provide the conclusion that a region represents information such that it can be classified. However, the answer of how remains to be determined. A simple remedy is to align the prediction scheme outlined above with the hypothesis of interest in what is termed *cross classification* in the context of this thesis. In an example of cross-classification, a classifier trained on objects at one retinal position and tested at another can be used to test whether visual object representations are position-tolerant (Cichy, Chen, & Haynes, 2011; Cichy, Heinzle, & Haynes, 2012; Kravitz, Kriegeskorte, & Baker, 2010). It has also been used to compare neural operations behind working memory and mental rotation (Albers, Kok, Toni, Dijkerman, & de Lange, 2013; Christophel, Cichy, Hebart, & Haynes, 2015). Using a classifier to show that representations are shared across domains allows us to invoke transitivity if we already understand processing in one domain, as the representation or computation must be shared across stimulus sets in order for the cross classification to be successful.

This method has potential in evaluating hypotheses concerning domain-generality. I apply cross-classification in chapter 2 to directly test the hypothesis that Broca's area acts as a supramodal structured sequence processor (Tettamanti & Weniger, 2006). By designing an experiment such that cognitive operations might be shared from one domain to the other, there are only two ways a classifier might be successful: 1) if it picks up on those common cognitive operations (assuming proper experimental controls) and 2) if noise structure is idiosyncratically different from domain to domain. Generally, we assume arbitrary noise will be constant within the same scanning session, especially if trials from different domains are

interleaved, and we can assume that arbitrary noise differences are actually meaningful signal. It is still important to consider previously discussed nature of successful classifications, particularly the pitfalls of interpreting classification accuracy, though relative comparisons are valid (Hebart & Baker, 2017).

While Chapter 2 demonstrates that cross-classification can be used to establish cross-domain representations, this approach is not appropriate for within-domain studies. For more subtle representations, or within-domain inferences, there is a renewed push for using representational similarity and encoding models (Naselaris et al., 2011; Diedrichsen & Kriegeskorte, 2017).

1.2.2 RSA

An increasingly popular alternative to MVPC is known as Representational Similarity Analysis (RSA- Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013; Nili et al., 2014). RSA characterizes the representation in a brain region by a representational dissimilarity matrix (RDM), and compares the empirical matrix with a model. An RDM is a square symmetric matrix, with each entry referring to the dissimilarity between the activity patterns associated with two trials (e.g., entry (1,2) would represent the dissimilarity between activity patterns of trial 1 and trial 2 for a given participant). Procedurally, each element of the RDM is calculated as 1 minus the Pearson correlation between the beta-series for each pair of trials. Hypothesis models are manually generated to reflect idealized RDMs expected if the group of voxels is indeed modulating its activity with respect to the manipulation (See chapters 3 and 4 for applications to higher cognition).

RSA addresses the information representation issue raised by MVPC at the cost of computational sophistication. By comparing distance matrices from several different models with subject performance (i.e. "second-order distance", Nili et al., 2014), it becomes possible to arbitrate between different representational models, thus providing insight as to the computational level operations implemented in the brain (Marr, 1982). For example, it has been used to compare deep learning architectures with different ROIs in the visual system (Seyed-Mahdi & Kriegeskorte, 2014). In chapter 3 we employ RSA to arbitrate between models of number representation in the parietal cortex, and in chapter 4 we use RSA to determine which regions of the brain modulate activity with the predictions of computational models of analogy.

While RSA doesn't suffer from rank-deficiency like MVPC analyses, it is typically applied in the ROI (Kriegeskorte et al., 2008) or searchlight (Kriegeskorte et al., 2006) approach in order to reduce the number of features. This is because in its basic formulation, it assumes all features (i.e. voxels) are of equal importance when calculating trial-by-trial distance (though see Kriegeskorte & Diedrichsen, 2016, Walther et al., 2016, and Diedrichsen & Kriegeskorte, 2017 for a recent improvement). By introducing a priori ROIs or limiting region size via searchlight, we maximize the chance that voxels meaningfully encode information and that the calculated distance metrics are valid and not driven by uninformative voxels. RSA also enjoys the advantage of being able to employ all samples (trials) within the dataset, rather than needing to partition as in MVPC (Diedrichsen & Kriegeskorte, 2017).

CHAPTER 2

The language of music: Common neural codes for structured sequences in music and natural language

2.1 Abstract

The ability to process structured sequences is a central feature of natural language but also characterizes many other domains of human cognition. In this fMRI study, we measured brain metabolic response in musicians as they generated structured and non-structured sequences in language and music. We employed a univariate and multivariate cross-classification approach to provide evidence that a common neural code underlies the production of structured sequences across the two domains. Crucially, the common substrate includes Broca's area, a region well known for processing structured sequences in language. These findings have several implications. First, they directly support the hypothesis that language and music share syntactic integration mechanisms. Second, they show that Broca's area is capable of operating supramodally across these two domains. Finally, these results dismiss the recent hypothesis that domain general processes of neighboring neural substrates explain the previously observed "overlap" between neuroimaging activations across the two domains.

2.2 Introduction

A central intuition in the study of human language as a cognitive phenomenon is the idea that, while listening to a linear signal such as speech, our minds spontaneously build abstract and structured hypotheses representing how discrete elements within a sequence relate to each other (Chomsky, 1957, 1983; Lashley, 1951; Fitch & Martins, 2014; Jackendoff, 2002; Monti, 2017). The use of such representations is most clearly displayed in natural language (Ding, Melloni, Zhang, Tian, & Poeppel, 2016; Berwick, Friederici, Chomsky, & Bolhuis, 2013), but also characterizes other aspects of human cognition, such as logical reasoning (Osherson, 1975; Monti & Osherson, 2012), algebraic cognition (Varley, Klessinger, Romanowski, & Siegal, 2005; Monti, Parsons, & Osherson, 2012; Maruyama, Pallier, Jobert, Sigman, & Dehaene, 2012), and music cognition (Patel, 2003; Katz & Pesetsky, 2011; Lerdahl, 2001), among others. The relationship between the syntactic operation of language and the syntax-like operations of other aspects of human cognition has thus been at the center of a long-standing debate concerning the degree to which human thought is embedded within, or enabled by, natural language (e.g., Lashley, 1951; Boeckx, 2010; Gleitman & Papafragou, 2013; Fitch & Martins, 2014; Fitch, 2014; Monti, 2017).

Lashley (1951) commented on the prevalence of structured sequences across domains, noticing that they exhibited the following three properties: (1) connectedness; i.e. no node is isolated from the others, (2) a root element; i.e. "sentence" or "chord" that is superior to others and (3) acyclic structure; establishing order as a unique property (Lashley, 1951; Fitch & Martins, 2014). In the context of music cognition, the analogy with the structural aspects of language is particularly pronounced. As discussed elsewhere (e.g., Lerdahl &

Jackendoff, 1985; Patel, 2003; Fadiga, Craighero, & Alessandro, 2009; Fitch, 2014; Peretz, Vuvan, Lagrois, & Armony, 2015), music and language are both characterized by discrete elements (e.g., words, chords) which can be (recursively) combined, according to specific rules, to form organized structures (e.g., sentences, melodies) which are typically encoded within linear, time-dependent, signals.

Nonetheless, whether this analogy is substantial or merely superficial remains a debated issue (cf., Peretz et al., 2015). At one end of the spectrum, it has been proposed that language and music are governed by the very same syntactic processes applied to different building blocks (e.g., words vs. notes). According to this view, "[a]ll formal differences between language and music are a consequence of differences in their fundamental building blocks[; i]n all other respects, language and music are identical" (Katz & Pesetsky, 2011). Along similar lines, it has been proposed that the common representations underlying the structure processing in language and music can be localized to the neural mechanisms encapsulated within the left inferior frontal gyrus (IFG; often referred to as Broca's Area), a region hypothesized to operate as a "supramodal hierarchical parser" (Tettamanti & Weniger, 2006; Fadiga et al., 2009). Consistent with this view, a rapidly growing neuroimaging literature has shown music processing to recruit cortical regions overlapping with areas known to be involved in syntactic and semantic aspects of natural language processing (Patel, Gibson, Ratner, Besson, & of cognitive, 1998; Maess, Koelsch, Gunter, & Friederici, 2001; Koelsch, Gunter, Cramon, & Neuroimage, 2002; Tillmann, Janata, & Research, 2003; Koelsch, Kasper, Sammler, & Nature …, 2004; Koelsch, Gunter, Wittfoth, & Sammler, 2005; Brown, Martinez, & Parsons, 2006), see (Rogalsky, Rong, Saberi, & Hickok, 2011), for a conflicting result). Nonetheless,

while the observation of overlapping neural substrates is often taken to imply the presence of shared neurocognitive representations between language and music, this is not necessarily the case (Peretz et al., 2015)) and indeed has never been shown to be true. This "missing link" in the neuroscientific literature leaves open the possibility that commonly recruited areas of the brain might, in fact, represent very different operations that do not translate, or align, across the two domains, or that are entirely unrelated to the processing of these relationships. In line with this observation, it has been suggested that language and music are in fact better thought of as modular and largely independent of each other (Marin & Perry, 1999; Peretz & Coltheart, 2003). In support of this view, a rich neuropsychological literature has described cases of individuals who exhibit amusia in the absence of aphasia, as well as aphasia in the absence of amusia (Luria, Tsvetkova, & of the neurological, 1965; Peretz, 1993; Peretz et al., 1994; Ayotte, Peretz, Rousseau, Bard, & Bojanowski, 2000; Piccirilli, Sciarma, & Luzzi, 2000; Ayotte, Peretz, & Brain, 2002).

The reason for the contradicting evidence is still a matter of debate. According to some, the fracture between neuropsychological and neuroimaging findings can be reconciled with a middle-ground solution in which language and music are viewed as partially overlapping systems (Patel, 2003; Patel, Iversen, Wassenaar, & Aphasiology, 2008)). Under this view, referred to as the shared syntactic integration resource hypothesis, language and music are characterized by both domain-specific (i.e., separate) and shared processes. The domain-specific processes relate to the particular features of each syntax, which are recognized as architecturally different, while shared processes provide neural resources for the activation of the relevant stored syntactic representations (Patel, 2012). According to others, the

inconsistency between the two sets of findings might instead be due to experimental and neuroanatomical considerations (Fedorenko & Varley, 2016). Specifically, the overlap often reported, in neuroimaging studies, in left inferior frontal regions could be a reflection of task-general demands tied to the use of structural-violation paradigms (e.g., the P600 and the early left/right anterior negativity effects reported in electrophysiological studies; Janata, 1995; Maess et al., 2001; Koelsch et al., 2002, 2005; Steinbeis & Koelsch, 2008; Tillmann et al., 2003; and later localized to the inferior frontal gyri through neuroimaging; Musso et al., 2015; Kunert, Willems, Casasanto, Patel, & Hagoort, 2015). Deviant events are indeed likely to elicit ancillary processes including attentional capture, detection of violated expectations, or error correction, regardless of whether the violation applies to natural language, music, arithmetic, or motor sequences. Such processes are unrelated to the extracting or forging of structured sequences and are known to elicit activation in domain-general regions (proximal or partially overlapping with Broca's Area; see Fedorenko & Varley, 2016, for a detailed discussion).

In the present study, we address the relationship between the mechanisms of natural language and those of music in a 3 Tesla functional magnetic resonance imaging (fMRI) within-subjects design in which competent musicians generate structures in language (active/passive voice sentences versus repeating a verb) and music (root/second-inversion position ascending triads versus repeating a note; cf., Figure 2.1 and Table 2.1). Crucially, we employ a (rarely explored) generation task to avoid the confound of salient events, and we use a multivariate cross-classification approach to resolve the interpretational ambiguity present in the previous neuroimaging literature (which has been specifically advocated for;

see Peretz et al., 2015), thereby helping resolve the question of whether natural language and music share a common underlying neural code for representing structured sequences.

2.3 Methods

Participants We recruited 21 total participants to reach the predetermined sample size (N=20, 8 female participants) based on previous literature Musso et al., 2015: N=11; Kunert et al., 2015: N=19; Koelsch et al., 2002: N=20). An additional subject was recruited because the data from one of the participants exhibited excessive motion during the procedure (see below). Participants received $50 compensation for taking part in the experiment. All participants were native English speakers, right handed, and competent musicians currently enrolled in the UCLA Herb Alpert School of Music. Participants were only enrolled if they could demonstrate proficiency in singing/generating both a root position and IInd inversion ascending triad arpeggio. Participants with perfect pitch were excluded. Participants signed informed consent prior to taking part in the session, as per the procedures approved by the UCLA Institutional Review Board.

Stimuli For both materials (i.e., "language" and "music" trials), the first cue was delivered visually, by presenting one of three icons in the middle of the screen. A ‘ ’ symbol indicated an active or root position trial (depending on whether the second cue was a word or a note, respectively); a ‘ ’ symbol indicated a passive or a IInd inversion trial; a ‘ ’ symbol indicated a non-structured (i.e., repeat) trial. The second cue was delivered aurally and consisted of either a verb or a note, thus revealing whether the trial was a language or music trial,

and allowing disambiguation of the instruction provided by the first cue. (See Table 2.1 and Figure 2.1 for sample stimuli.) Cues for language trials consisted of seven monosyllabic, reversible, present tense verbs (i.e., "bring," "tell," "teach," "throw," "leave," "give," "pay"). Cues for the music trials consisted of 7 notes recorded with an electronic piano. The notes comprised a chromatic scale (i.e. each consecutive pitch was separated from the following by a semitone, or a half step). Participants were allowed to choose, among three sets of musical cues (high, middle, or low), whichever pitch range most comfortably matched their vocal range. For each set, a perfect fifth (7 semitones/half steps) separated the lowest and highest pitches of the cues. Combining the 7 cues (i.e., words/notes), 3 generative rules (2 structured, 1 repeat), and 2 materials (i.e., language, music), resulted in 42 unique trials (i.e., 21 per material type).

Experimental Design Each participant viewed the 42 unique trials twice (once in the first two runs, once in the second two runs). Trial types were equally distributed across 4 runs, and, within each, randomly presented. Stimuli were presented using PsychoPy (Peirce, 2009); visual cues were displayed through a custom-made MRI-compatible projection system while auditory cues were delivered through a Magnetic Resonance headphone system. As shown in Figure 2.1, each trial started with the generative rule cue (i.e., '♢, ♣, ↻'), displayed on screen for 1.5 s, followed by the second cue (i.e., word or note) presented, aurally, for 1.8 s. After a variable jitter (between 6 and 8 s), a fixation symbol blinked four times (with a cycle of 0.8 s of display and 0.35 s interval). The first blink (with a black square symbol) served as a warning that the "performance/response" period was to begin. The following three blinks (with a black circle symbol) marked the performance/response period and provided a tempo

Language	**Active**	**Passive**	**Repeat**
Cue 1	♢	♣	↺
Cue 2	"Give"	"G"	"Tell"
Response	"X gave Y"	"X was given Y"	"tell, tell, tell"
Music	**Root position**	**IInd Inversion**	**Repeat**
Cue 1	♢	♣	↺
Cue 2	"C"	"D"	"E"
Response	"C-E-G"	"D-G-B"	'E-E-E"

Table 2.1: Sample cues & stimuli

for responding. The tempo was never varied, neither within nor across subjects, and was only employed to provide participants with a consistent rhythm for responding.

Finally, a variable length fixation screen (with a random jitter between 5 and 7 s chosen, on a trial-by-trial basis, from an exponential distribution) separated each trial from the subsequent one. Each run lasted, on average, 293.57 s (S.D. = 15.81). Participants were trained to asymptotic performance prior to the imaging session, in a separate room, after having signed informed consent. The experimenter corrected any errors the participant made until satisfactory performance was achieved (less than 2 errors per block of trials). Training ceased when participants could perform at least 12 out of 13 trials correctly, minimizing the sound production time across conditions.

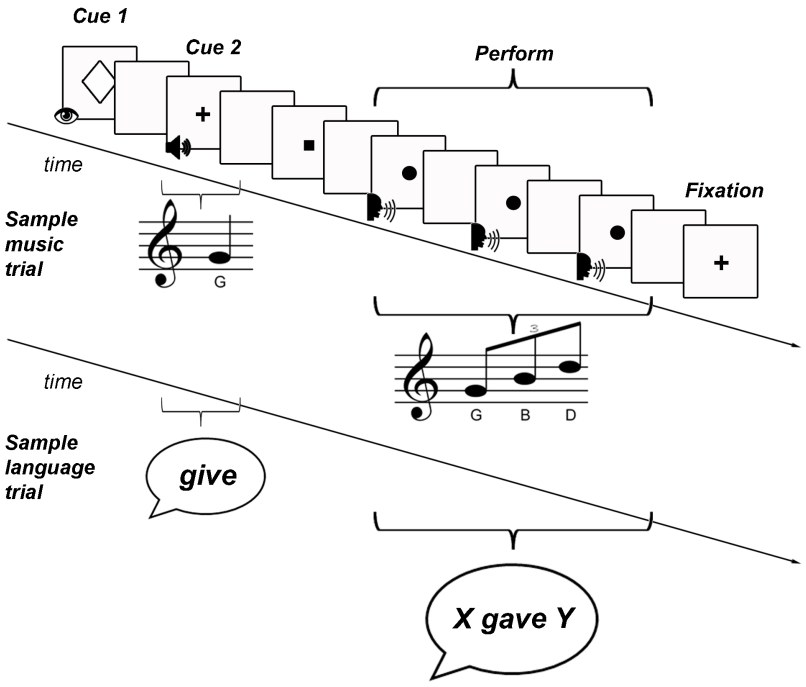

Figure 2.1: Sample music and language trials timelines.

Data Acquisition Data were acquired on a 3 Tesla Siemens Tim Trio Magnetic Resonance Imaging (MRI) scanner at the Staglin IMHRO Center for Cognitive Neuroscience at UCLA. Structural data were acquired using a T1-weighted sequence (MP RAGE, TR = 1,900 ms, TE = 2.26 ms, voxel size 1 mm3 isovoxel). Blood oxygenation level dependent (BOLD) data were acquired with a T2*-weighted Gradient Recall Echo sequence (TR = 3,000 ms, TE = 35 ms, 45 interleaved slices, voxel size $3 \times 3 \times 3.3$ mm) with prospective motion correction in order to reduce the impact of subject motion during performance.

Data Preprocessing Data analysis was carried out using FSL (Smith et al., 2004). Prior to analysis, data underwent a series of conventional preprocessing steps including motion correction, slice-timing correction (using Fourier-space time-series phase-shifting), spatial

smoothing using a Gaussian kernel of 5 mm full-width half-max, and highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with =50.0s). Data from each individual run were analyzed employing a univariate general linear model approach (Monti, 2011) inclusive of a pre-whitening correction for autocorrelation. Following current convention, any participant exhibiting average motion greater than 3 mm was excluded (N=1).

Univariate Analysis For each run of each participant, a univariate analysis was conducted using, as the main variables of interest, 6 regressors, one per trial type (i.e., language active voice, language passive voice, language repeat, music root position, music IInd inversion position, music repeat). Regressors marked the performance/response period of each trial (see Figure 2.1). A number of additional nuisance regressors were employed to model cue periods, motion (including first and second derivatives, and their difference), as well as the short intervals between the second cue and task performance. This last regressor is particularly important since it parcels out periods in which subjects are likely to be engaging in strategies in anticipation of the task, which, in the absence of any participant feedback, are un-controlled and thus difficult to interpret. For each run we computed 4 contrasts: structured versus repeat trials for language and music materials ("simple effect" contrasts), separately, and the interaction between the two simple effects ("interaction contrasts") in both directions (i.e., simple effect of structured trials in language greater than the simple effect of structured trials in music, and vice versa). Data from individual runs were aggregated employing a mixed effects model (i.e., employing both the within- and between-subject variance), and using automatic outlier detection. Z (Gaussianised T) statistic images were

thresholded using a cluster correction of Z>2.3 and a (corrected) cluster significance threshold of P=0.05.

Multivariate Analysis The input to the multivariate analysis was a set of volumes of regression coefficients (i.e., "$\hat{\beta}$") marking the magnitude of activation, for each voxel, in each trial (per subject). These trial-wise "patterns of activations" were obtained by employing the iterative Least Squares – Separate approach (LS-S; Mumford et al., 2012) in which a separate GLM is run (here, using FILM with local autocorrelation) for each trial. At each iteration, one regressor marks the trial of interest, while all remaining trials are collapsed into a nuisance regressor (see (Mumford et al., 2012), Figure 2.1 for a visual depiction of this approach). This approach has been shown, in simulations, to produce activation estimates that have the highest correlation with true activation magnitudes (Mumford et al., 2012), and has also been shown to adapt best to multivariate analyses when used in conjunction with full randomization of trials (different for each subject) and with equal inter stimulus interval across condition (Mumford et al., 2012), as we have done. The patterns of activation were then concatenated across time to construct a subjectwise "$\hat{\beta}$-series" of activation magnitude per trial per voxel (Rissman et al., 2004).

In order to assess whether natural language and music share underlying neural representations, we employed a cross-classification searchlight analysis using a linear support vector machine (SVM) algorithm. Cross-classification was performed by training the SVM classifier to recognize structure vs. repeat trials in one domain, and then attempting to classify structure vs. repeat structure trials in the other domain ("M2L" and "L2M" cross-classifications for training on music and testing on language and vice versa, respectively). L2M cross-

classifications were performed over voxels found significant in the structure minus repeat trials for language materials (only); M2L classifications were performed over voxels found significant in the structure minus repeat trials for music materials (only). Because the significant voxels in the two univariate contrasts could overlap, this feature selection ensures that the training and testing datasets for each type of cross-classification (i.e., L2M, M2L) remain completely separate, thereby avoiding any bias in the analysis. Classifications were performed on a single subject basis, in native space, employing a 6 mm radius searchlight approach (Kriegeskorte et al., 2006). To account for the imbalance between the number of structure and repeat trials (28 and 14, respectively, per each domain) and avoid biasing the classifier, we performed a resampling procedure in which, at each of 1000 iterations, a subsample of 14 (structured) trials was randomly selected, in order to train and test the classifier on a matching number of trials across conditions. Results across the 1000 iterations were averaged to yield a single classification accuracy value for each searchlight sphere.

Then, in order to assess whether the two structured sequences of each domain could be distinguished from one another on the basis of patterns of brain activity within Broca's area (i.e., Language Active Voice (L_{AV}) versus Language Passive Voice (L_{PV}) and Music Root Position (M_{RP}) versus Music 2nd Inversion (M_{SI})), we performed a second searchlight analysis (Kriegeskorte et al., 2006). Specifically, for each domain separately, we performed an SVM classification on a single subject basis, in native space, employing a 6 mm radius searchlight. At each iteration of the searchlight, classification accuracy was assessed using a leave-one-run-out cross validation procedure in which the SVM was trained on the trials from 3 runs, and accuracy was assessed on the trials from the remaining run. This procedure was

repeated for each run, and the 4 accuracies were then averaged to obtain overall accuracy for each classification.

For both classification analyses, statistical significance was assessed, at the group level, employing a permutation-based sign test and against a criterion of p= 0.05 corrected for multiple comparisons at the cluster level (using FSL's threshold free cluster enhancement, Winkler, Ridgway, Webster, Smith, & Nichols, 2014). At the single subject level, significance was assessed with a permutation procedure, in which the classification was repeated 1000 times, with shuffled testing labels, to construct a null distribution for each voxel (cf., Etzel & Braver, 2013). Classifications falling within the top 5% of the null distribution were considered significant.

2.4 Results

Univariate Analysis The simple effect contrast of structure versus repeat trials for language materials uncovered a set of expected activations in left inferior frontal gyrus (including its pars opercularis and triangularis, in Brodmann Areas [BA] 44, 45), posterior middle and superior temporal cortices (BA 21, 22), bilateral parietal (spanning BA 7, 40) and medial (BA 6), middle (BA 8), and superior frontal (BA 6) areas (mostly left lateralized; see Figure 2.2, below, and Table A.1 in the Supplemental Material available online for complete list of local maxima).

When performed on music trials, the same contrast uncovered a number of activation clusters across bilateral frontal and parietal regions (see Figure 2.2 and Table A.2). The frontal cluster included bilateral maxima in the inferior frontal gyri (spanning its pars op-

ercularis in BA 44, triangularis in BA 45, and orbitalis in BA 47), rostral insular cortex (spanning BA 13 and its junction with 45 and 47), as well as bilateral foci across middle (BA 6), superior (BA 6, 8) frontal, and cingulate (BA 32) gyri. In addition, bilateral activations were observed in the inferior (BA 40) and superior (BA 7) parietal lobuli, as well as in the posterior cerebellum (see Table A.2 in the Supplemental Material available online for the complete list of local maxima). As shown in Figure 2.2 (regions in yellow), the structure versus repeat contrast uncovered a number of common areas across language and music materials, including the left inferior frontal (in its pars opercularis, BA 44) and middle frontal (in BA 6) gyri, as well as the medial frontal/cingulate gyri (BA 6, 32), and bilateral posterior parietal lobe (in both BA 7 and 40). In order to avoid interpreting a "reverse subtraction", we characterized the mean activity profile for structure and repeat conditions to identify the primary driver in IFG. Mean z-scores from the IFG subregions (defined by external atlases: pars opercularis and pars triangularis from Harvard-Oxford and pars orbitalis from AAL) are displayed in figure A.2.

The interaction of structure versus repeat structure and materials revealed the left superior and middle temporal gyri (BA 21, 22) to be specific to language (see blue areas in Figure A.1 and Table A.3), whereas foci surrounding the right orbital and sub-lobar segments of the inferior frontal gyrus (mainly in BA 47 and 13), along with right superior frontal (BA 6), medial frontal (BA 6, 32) and contralateral posterior cerebellum appeared to be specific to structure in music (see Figure A.1 and Table A.3 in the Supplemental Material available in A).

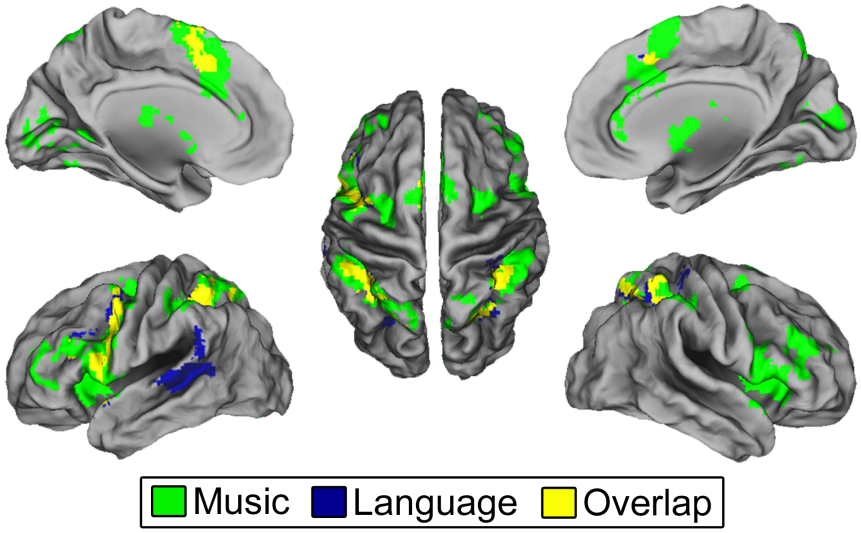

Figure 2.2: Overlay of the structure versus repeat contrast results for language (blue) and music (green) (yellow marks overlap between the two tasks).

Multivariate Analysis In order to assess whether natural language and music share neural codes for representing structure, we performed a cross-classification multivariate pattern analysis. In this approach, an SVM classifier was trained to recognize structure versus repeat trials in one domain and was then tested on the other. That is to say, we trained a classifier to distinguish structure versus repeat trials in language and then tested it by assessing its ability to discriminate structure versus repeat trials in music ("L2M" cross-classification), and vice versa ("M2L" cross-classification; see Methods).

As shown in Figure 2.3, significant cross-classifications were evident across a number of regions within medial prefrontal cortex, bilateral posterior parietal cortices, as well as left precentral, inferior (in the pars opercularis), and middle frontal gyri, matching areas of univariate overlap between the two domains (i.e., yellow regions in Figure 2.2). In addition, within each of these cross-classification clusters are areas (in yellow) capable of performing

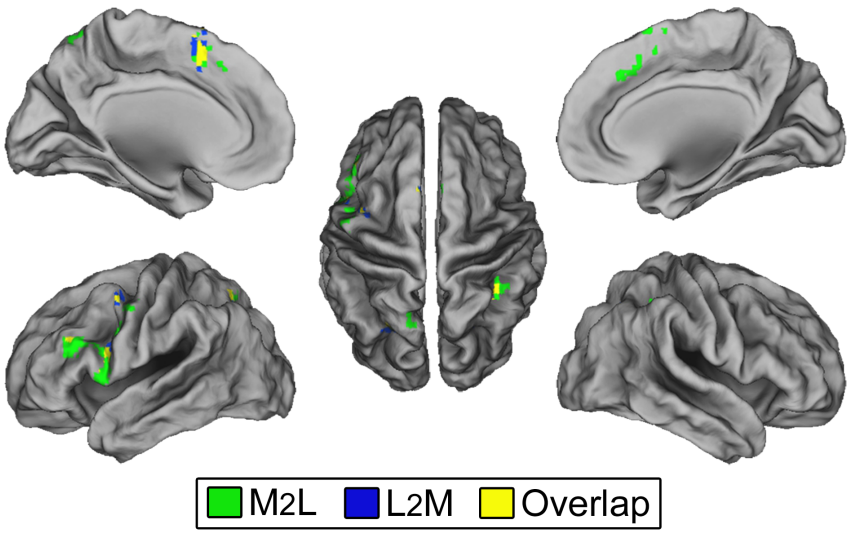

Figure 2.3: Green regions represent areas in which the SVM algorithm could significantly classify, with above chance accuracy, structure vs. repeat trials in language materials after having been trained to recognize structure vs. repeat trials in music materials (i.e., M2L cross-classification). Blue regions represent areas in which the SVM algorithm could significantly classify, with above chance accuracy, structure vs. repeat trials in music materials after having been trained to recognize structure vs. repeat trials in language materials (i.e., L2M cross-classifications). Yellow areas show searchlight centers that can significantly perform both classifications.

both L2M and M2L classifications, further demonstrating some extent of common underlying neural representation across the two domains. Crucially, this effect could be observed at the single-subject level, with median (single-subject) cross-classification accuracies at 61% for both L2M and M2L classifications, and ranges between 59% and 65%, and 58% and 64% for L2M and M2L classifications, respectively; with chance being 50%). Focusing on the left inferior frontal gyrus in particular, Figure 2.4 depicts the reliability of the result at the single-subject level. Figure 2.4 also demonstrates a significant across-subject variability in the exact location of voxels sensitive to linguistic structure within the inferior frontal gyrus (as previously shown; Fedorenko, Hsieh, A, S, & Kanwisher, 2010), something that we also observe in music-structure sensitive voxels, resulting in a systematic but variably located overlap in voxels capable of both L2M and M2L classifications within this region.

We then performed a searchlight classification in order to determine whether Broca's area was capable of discriminating between the two structured conditions (Active versus Passive voice in Language, and Root vs 2nd Inversion position in Music). After correcting for multiple comparisons, there was no evidence at the group level that regions able to classify the structures within domains were consistent across subjects. However, at the single subject level, a within-subject permutation test revealed that all 20 subjects exhibited, within Broca's area, regions capable of discriminating between the two language structures (L_{AV} v L_{PV}; cf. Figure 2.5, blue areas), 17 subjects exhibited regions capable of discriminating between the two music structures (M_{RP} v M_{SI}; cf. Figure 2.5, green areas), and for 8 subjects the two sites overlapped (cf. Figure 2.5, yellow areas).

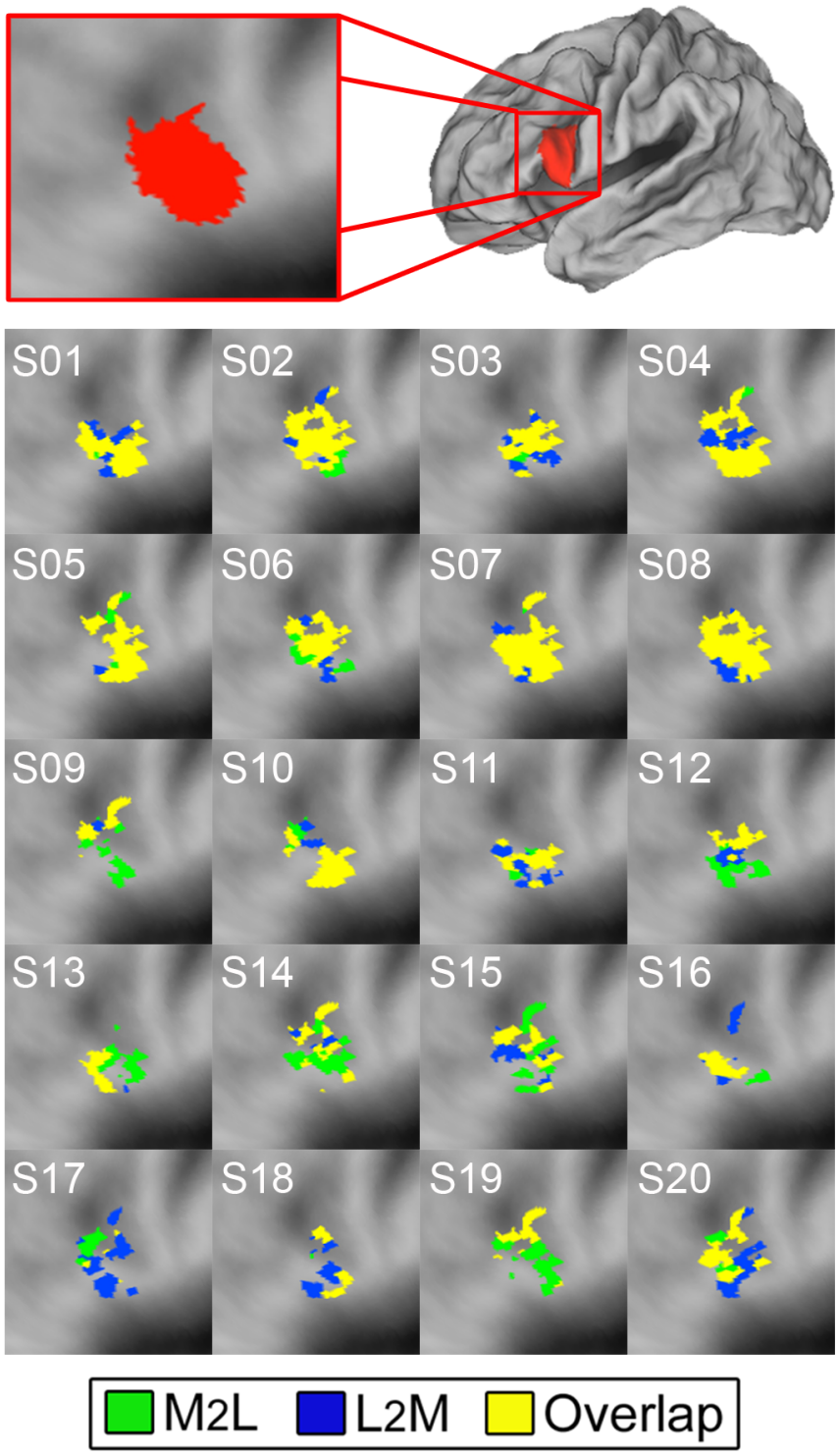

Figure 2.4: Cortical flat-maps depicting, for each participant separately, searchlight centers capable of significant cross-classifications (L2M in blue; M2L in green; overlap in yellow) within the inferior frontal gyrus as defined anatomically (highlighted in red, at the top). Each image (labeled as 'S##') represents the classification results for a single participant.

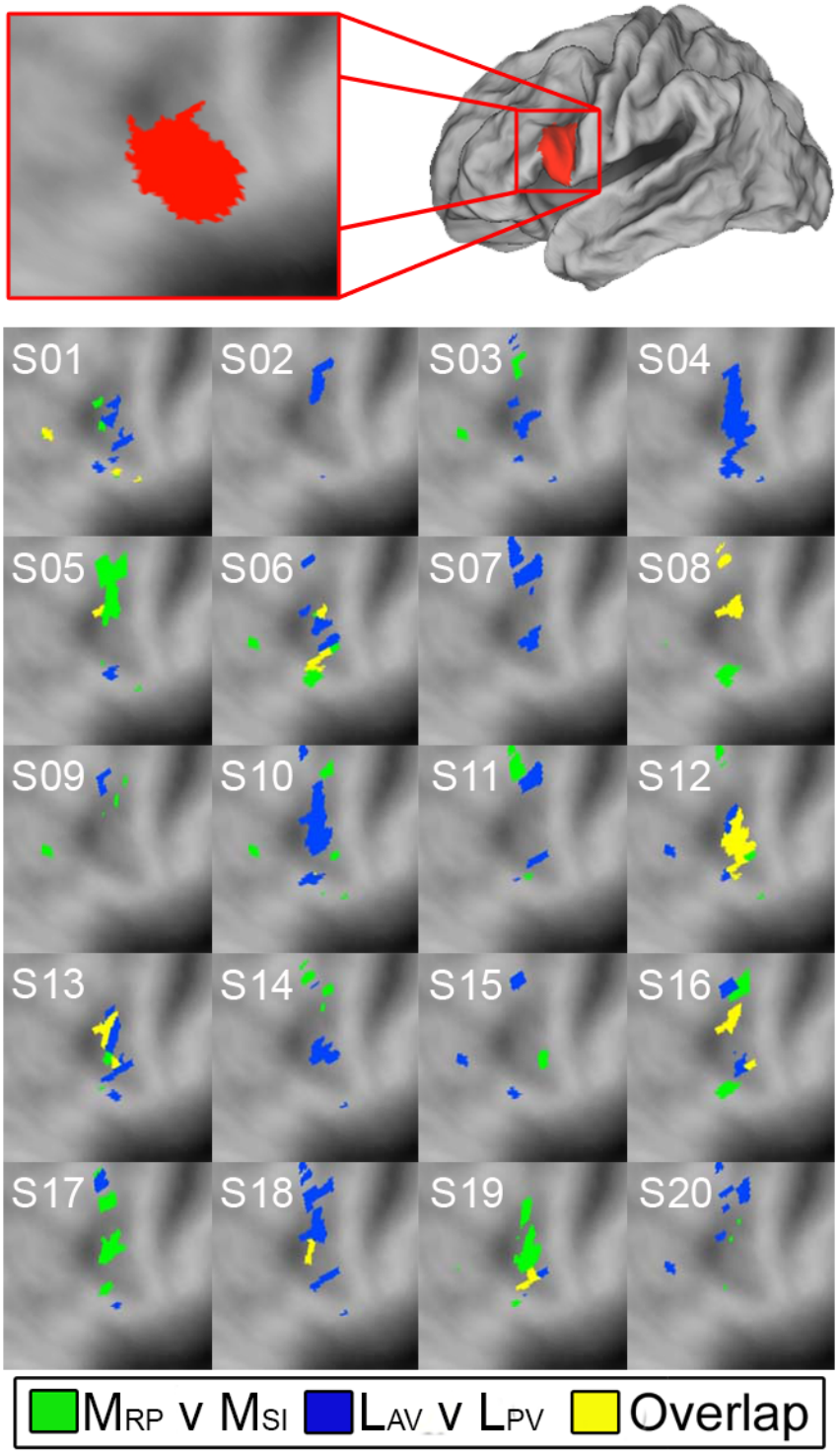

Figure 2.5: Cortical flat-maps depicting, for each subject, searchlight centers capable of discriminating between structures within each domain (L_{AV} v L_{PV} in blue, M_{RP} v M_{SI} in green, Overlap in yellow) within the inferior frontal gyrus.

2.5 Discussion

In this study we have addressed the question of the relationship between natural language and human cognition by contrasting the neural substrates accompanying the generation of structured sequences across language and music. Overall, our results provide direct evidence for the hypothesis that language and music have a shared neural code for producing structured relationships – a phenomenon that we observe both at the group as well as at the single subject level.

More specifically, we report three central findings. First, employing a magnitude-based univariate approach, we find the generation of structured sequences in language to recruit a well-known left lateralized network of frontal and temporal regions, along with posterior parietal foci, while generation of music sequences engaged a larger, and strongly bilateral, set of fronto-parietal regions. The neural substrate elicited by this performance paradigm (which has remained almost unexplored in the context of music, with the exception of Brown et al., 2006) matches very closely the neural substrate typically reported in tasks focusing on competence in both language (e.g., Michal, Hendler, Kahn, Dafna, & Grodzinsky, 2003; Monti, Parsons, & Osherson, 2009) and music (e.g., Maess et al., 2001; Koelsch et al., 2002, 2005; Cheung, Meyer, Friederici, & Koelsch, 2018.

Our second main finding, as evaluated with the same univariate activation-based approach, shows that the building of structured sequences in language and music relies on a number of common regions across left lateral and medial frontal cortices, as well as bilateral posterior parietal regions. In particular, the univariate analysis shows that the posterior-most aspect of Broca's Area, in the pars opercularis of the left inferior frontal gyrus, is

metabolically responsive to the presence of structure in the context of both language and music materials (Figure A.2) – a finding that is consistent with results from previous studies Maess et al., 2001; Koelsch et al., 2002, 2004, 2005; Brown et al., 2006). Beyond the left inferior frontal gyrus, our findings show that the interplay between language and music might well extend to a set of regions in frontal and parietal cortices conventionally referred to as the multiple-demands network (Duncan, 2010), which have been shown to be recruited across a broad class of cognitive operations (Fedorenko, Duncan, & Kanwisher, 2013), and medial frontal regions that are commonly recruited by motor planning (Tanji & Shima, 1994; Tanji, Shima, & Mushiake, 2007; Haggard, 2008); and that have also been shown to play a role in rhythm perception and generation of spontaneous movement in response to music stimuli (Grahn & Brett, 2007; Lima, Krishnan, & Scott, 2016). The absence of temporal regions identified by the domain-specific main effects (Figure A.1), specifically the involvement of posterior STS (i.e. "Wernicke's area") in language tasks but not music, corroborates previous work reporting that posterior temporal regions might engage a semantic/syntactic interaction (See Friederici, 2011, 2012 for a review).

Finally, our third, and crucial, finding addresses the significance of the frequently reported overlap between the neural substrate of language and that of music, thereby directly addressing the question of whether the mechanisms of natural language play a role in processing the structured sequences of music. Indeed, while regions of overlapping activation for these two domains have been widely interpreted as marking areas of shared neurocognitive processing (Kunert et al., 2015), these hypotheses had not been directly tested (until now), prompting some to specifically advocate multivariate analyses such as the one adopted

here (Peretz et al., 2015). As we reported above, we could find within each of the regions of univariate overlap (in Broca's area), areas capable of recognizing music structure on the basis of language structure and vice versa. In fact, in each of these areas a subset of voxels could perform, at the same time, cross-classifications in both directions (i.e., L2M and M2L), demonstrating a degree of shared neuronal representation of structures across domains. Furthermore, although the result was not significant at the group level, within Broca's area we could also find, at the single subject level, variably located sub-regions sensitive to the different structures of language (i.e., active versus passive voice) and music (i.e., root versus 2nd inversion position). These findings thus provide evidence in favor of the idea that language and music cognition share, within Broca's area, resources related to establishing structured relationships tying discrete elements into well-formed complex structures (Patel, 2003). Nonetheless, our data cannot address the issue of whether the neural representation of specific operations (e.g., syntactic movement, Chomsky, 1957, 1995) can be directly mapped across the two domains, as entailed by some views (Katz & Pesetsky, 2011).

In interpreting our results, however, two important limitations should be noted. First, it is not possible, from our analysis, to determine the precise factor(s) driving the successful classification of either structured versus non-structured trials across domains, or within the structured sequences of each domain (cf., Hebart & Baker, 2017). Indeed, in both analyses, stimuli differed across a number of important features including the presence of structured sequences as well as working-memory demands (as shown by the common recruitment of regions within the so-called multiple demands (Duncan, 2010), or extrinsic mode (Hugdahl, Raichle, Mitra, & Specht, 2015), network). The degree to which each of these (and possibly

other) factors drives our results cannot be assessed with our own data. Nonetheless, while the co-localization of cross-classifying voxels within Broca's area is open to such interpretational ambiguity, it is harder to make the same case for the little co-localization observed in the classification of different ordered structures (within each domain). If indeed general processes were responsible for the activations previously observed for music in Broca's area (Fedorenko & Varley, 2016)), a greater co-localization of voxels capable of classifying the two structures within each domain would be expected. Furthermore, a recent study has shown that increased (general) working memory demands are not sufficient to drive activation within Broca's area, while activating a well-characterized network of frontal and parietal regions (Coetzee & Monti, 2018).

A second important consideration is the fact that, for both music and language, our structures were relatively simple as compared to previous work attempting to relate linguistic and music sequences (e.g., Kunert et al., 2015). It is thus not clear whether the fact that we fail to uncover a group effect in classifying, within each domain, the two structures (i.e., active/passive voice and root/2nd inversion position, for language and music respectively) is due to this difference or other factors such as the high variability of the fine distribution of neural representations across individuals (cf., Figure 2.5; see Fedorenko et al., 2010).

Overall, our findings are consistent with the idea that Broca's area might play a role across multiple domains, in the context of processing structured sequences, as envisioned in the Supramodal Hierarchical Parser hypothesis (cf., Tettamanti & Weniger, 2006). Nonetheless, it is important to note that to date this hypothesis has only found support in a narrow sense (e.g., as conceived in Fadiga et al., 2009, and see de Cavey & Hartsuiker, 2016 for

evidence of domain-general mechanisms), as it does not appear to extend to the hierarchical relationships of algebra (e.g., Varley et al., 2005; Monti et al., 2012), logic inference (e.g., Monti, Osherson, Martinez, & Parsons, 2007; Monti et al., 2009; Monti & Osherson, 2012), and spatial cognition (e.g., Bek, Blades, Siegal, & Varley, 2010). Thus far, it has only been found to be consistent with findings in the domain of language (cf., Bookheimer, 2002), music (here and in most previous neuroimaging reports; e.g., Maess et al., 2001; Koelsch et al., 2002; Koelsch, Rohrmeier, Torrecuso, & Jentschke, 2013), and motor actifon sequences (e.g., P. Fazio et al., 2009). In this sense, Broca's area cannot be viewed as a central parser capable of operating in any domain of cognition as one would expect of a domain general processor or working memory component, though it may be a core component in a network of regions engaging in hierarchical processing (Fitch, 2014)). While it might be speculated that Broca's area plays a role in cognitive domains where structured relationships trigger automatic and effortless intuitions (compare the ease of detecting a non-grammatical sentence or a sour note with the much more taxing task of detecting an incorrect algebraic expression or logic argument), the crucial factor(s) determining its involvement in processing structured sequences remains to be fully specified.

In conclusion, this report provides direct evidence that forging the structured sequences of natural language and music relies also on a common neural representational space which includes Broca's area, a region traditionally associated with the syntactic operations of language. Contrary to recent discussion (Fedorenko & Varley, 2016), we find that activations within Broca's area in response to music stimuli cannot be discounted as epiphenomenal to salient events (e.g., violations, which were not present in our design) or due to recruitment

of domain general processes (encapsulated in, or close to, Broca's area). Whether the common mechanisms originally evolved in connection with one of the two domains, or whether they jointly inherited their properties from a common anteceding cognitive domain (e.g., a "prosodic protolanguage," (Fitch, Hauser, & Chomsky, 2005); or the capacity for hierarchical sequences of motor actions, (Lashley, 1951), remains to be understood.

CHAPTER 3

Neural representations of magnitude for natural and rational numbers

3.1 Abstract

Humans have developed multiple symbolic representations for numbers, including natural numbers (positive integers) as well as rational numbers (both fractions and decimals). Despite a considerable body of behavioral and neuroimaging research, it is currently unknown whether different notations map onto a single, fully abstract, magnitude code, or whether separate representations exist for specific number types (e.g., natural versus rational) or number representations (e.g., base-10 versus fractions). We address this question by comparing brain metabolic response during a magnitude comparison task involving (on different trials) integers, decimals, and fractions. Univariate and multivariate analyses revealed that the strength and pattern of activation for fractions differed systematically, within the intraparietal sulcus, from that of both decimals and integers, while the latter two number representations appeared virtually indistinguishable. These results demonstrate that the two major notations formats for rational numbers, fractions and decimals, evoke distinct neural representations of magnitude, with decimals representations being more closely linked to

those of integers than to those of magnitude-equivalent fractions. Our findings thus suggest that number representation (base-10 versus fractions) is an important organizational principle for the neural substrate underlying mathematical cognition.

3.2 Introduction

Representations of Symbolic Number Types Humans are unique in having developed symbolic notations for numbers. Given that a primary function of numbers is to convey magnitude values, it is important to understand the mental and neural representations of numerical magnitudes. The goal of the current study was to address the question of how different symbolic notations (natural numbers, fractions, and decimals) map onto magnitude codes. Specifically, we sought to determine whether different notations map onto a single, fully abstract, magnitude code, or whether separate representations exist for specific number types (e.g., natural versus rational) or number representations (e.g., base-10 versus fractions). Numerous studies of numerical magnitude comparisons have yielded a symbolic distance effect: comparisons of numbers that are closer in magnitude (e.g., 7 vs. 8) are slower and more error prone than comparisons of numbers that are farther apart (e.g., 2 vs. 8; Moyer & Landauer, 1967; Holyoak, 1978). A similar distance effect is observed in children (Barth, La Mont, Lipton, & Spelke, 2005; Brannon, 2002). Rhesus monkeys display a distance effect for numerosity comparisons; moreover, they are capable of learning shapes (Arabic numerals) corresponding to small numerosities (1-4 dots), such that the shapes acquire neural representations overlapping those of the corresponding perceptual numerosities (Diester & Nieder, 2007).

The distance effect and other phenomena have been interpreted as indications that numerical magnitudes (at least for integers) are accessed quite automatically, and are associated with an analog magnitude representation akin to a mental number line (Dehaene & Changeux, 1993; Gallistel, 1993; Opfer & Siegler, 2012). Neuroimaging studies with both adults and children have implicated the intraparietal sulcus (IPS) as the central area for representing and comparing symbolic integer magnitudes (and also non-symbolic magnitudes) (Dehaene, Piazza, Pinel, & Cohen, 2003; Nieder & Dehaene, 2009; Piazza, Pinel, Le Bihan, & Dehaene, 2007; Pinel, Dehaene, Riviere, & LeBihan, 2001). Further, IPS activation is inversely related to the numerical distance between two numbers being compared (Cohen, 1973; Kaufmann, Koppelstaetter, Delazer, & Siedentopf, 2005), consistent with the behavioral distance effect.

While the representation of whole-number magnitude has received considerable attention, far less is known about the representation of other symbolic number types, such as the rational numbers (fractions and decimals). Some have argued that the representation of magnitude in general is entirely abstract, and that all symbolic and non-symbolic magnitudes can be represented using a single mental (and neural) number line (Eger, Sterzer, Russ, Giraud, & Kleinschmidt, 2003; Naccache & Dehaene, 2001; Siegler, Thompson, & Schneider, 2011). However, studies investigating this topic have as yet failed to reach a consensus. Previous behavioral research has mainly focused on the extent to which fractions are represented holistically. This work has focused on the issue of whether the overall magnitude of a fraction is accessed automatically, like an integer (Kallai & of Human, 2009; Meert, Grégoire, & Noël, 2010; Meert, Grégoire, & of Psychology, 2010; Schneider & Siegler,

2010; Sprute & Temple, 2011). Evidence for automatic and holistic magnitude representation come from studies examining the distance effect during fraction comparisons. Many studies (e.g. Schneider & Siegler, 2010) have found that adults show a distance effect when representing fractions during comparisons. However, other studies have shown that depending on the stimuli and availability of various shortcut strategies, adults may represent only the whole-number components of the fraction and not its holistic magnitude (e.g. Bonato, Fabbri, Umiltà, & Zorzi, 2007; L. K. Fazio, DeWolf, & Siegler, 2016).

Moreover, other work has shown that even when a distance effect is found for fraction comparisons, the size and scale of the effect is entirely different for fractions relative to either integers or decimals. DeWolf, Grounds, Bassok, & Holyoak, 2014 had adults compare fractions, matched decimals (rounded to three digits) and integers (created by multiplying the equivalent decimal by 1000 to obtain a three-digit integer). Comparisons for all three number types yielded reliable distance effects, based on the holistic magnitudes of the numbers being compared. Importantly, however, response times and error rates for the fraction comparisons were much higher than for comparisons of either decimals or integers, with the latter number types showing no differences in response times or errors. Moreover, the distance effect was much more pronounced for fractions, with response times averaging between 2 and 8 seconds for far versus near number pairs. In contrast, response times for integers and decimals overlapped with one another, and generally were no longer than 2 seconds. This dramatic difference in the scale of the distance effect across number types suggests that the magnitude information associated with fractions may be less precise than that associated with integers or decimals, and that the process of accessing magnitudes is more effortful and less automatic

for fractions than for either integer or decimal formats.

Using fMRI to Investigate Magnitude Representation Behavioral research investigating rational number magnitudes suggests there are important differences between magnitude processing for fractions relative to other number types. Although neuroimaging methods, and fMRI in particular, have been employed to assess the neural substrates of numerical magnitude representation (e.g., Damarla & Just, 2013), numerical symbols representations (see Ansari, 2007) and algebra (e.g., Monti et al., 2012), there is no consensus regarding the interpretation of the behavioral differences observed between fractions and other number types. The present study applied neuroimaging methods to assess the relationships among the neural representations of magnitude for different symbolic formats. If the representation of magnitude is entirely abstract, then the neural representations of a fraction and its magnitude-equivalent decimal (e.g., 2/5 vs. .40) in the IPS might be expected to be identical. In contrast, if fractions and decimals are processed very differently (as some behavioral studies suggest), then the neural codes for the different notations may differ. To date, these alternative predictions remain untested. In fact, only two studies have ever probed the neural representations underlying the processing of fractional numbers (Ischebeck, Schocke, & Delazer, 2009; Jacob & Nieder, 2009a), and neither of these assessed the neural representations underlying decimal numbers, or the relationship between neural representations of magnitude across different formats for rational numbers.

A few other studies have examined how neural representations of magnitude differ as a function of notation by comparing neural responses to whole numbers versus their verbal equivalents (e.g., "12" versus "twelve"). Some studies have found that IPS activation was

notation-independent (Eger et al., 2003; Naccache & Dehaene, 2001), whereas other studies suggest there may be both notation-specific and notation-independent areas (Bulthé, De Smedt, & Op de Beeck, 2015; Cohen Kadosh, Cohen Kadosh, Kaas, Henik, & Goebel, 2007; Damarla & Just, 2013). However, these studies all compared a single mathematical notation (whole numbers) versus natural language (number names). No work has been done to investigate the question of whether alternative mathematical formats, such as fractions versus decimals, evoke similar or distinct neural representations of magnitude.

As noted above, only two studies have investigated the representation of symbolic fraction magnitudes using functional magnetic resonance imaging (fMRI). Jacob & Nieder, 2009a used an adaptation paradigm to test symbolic fraction magnitudes (single and multi-digit fractions). Recovery in the BOLD signal after habituation was observed in the frontoparietal cortex, and specifically the IPS. The pattern of signal recovery was the same after presentation of either a new symbolic fraction (e.g., "1/2") or a new fraction written as a word (e.g., "half"), suggesting that fractions and their verbal equivalents recruit the same or overlapping neural areas.

The second study that investigated symbolic fraction notation with fMRI used a magnitude comparison paradigm, rather than an adaptation paradigm. Ischebeck et al., 2009 had adult participants perform a simple magnitude comparison task with fractions, in which participants saw two fractions simultaneously on the screen and pressed a button to indicate which was larger in numerical magnitude. The stimuli included different types of fraction pairs, some with common components, in order to enable a variety of potential strategies during the comparison process. The results showed that activity in the right IPS was inversely

correlated with the distance between the two fractions based on their holistic magnitude difference, and not with the distances between any component parts. Ischebeck et al. interpreted their fMRI results as supporting the hypothesis that (despite an opportunity to use componential strategies) fraction comparisons were performed using holistic magnitudes.

However, neither Ischebeck et al., 2009 nor Jacob & Nieder, 2009a directly compared processing of fractions with that of other symbolic formats. Although previous work indicates that magnitude representations for fractions involve roughly the same general neural area (the IPS) as do magnitude representations for symbolic integers (and non-symbolic numerosities; see Jacob & Nieder, 2009b; Jacob, Vallentin, & Nieder, 2012), the extent to which processing and representation of magnitude is the same or different for fractions relative to other number types has not been examined. Furthermore, the more general question of whether different symbolic formats for numbers evoke the same or different abstract magnitude representations remains unanswered.

The Present Study In the present experiment, we employ univariate and multivariate analysis of fMRI data to compare, in a within-subject design, the neural representations of magnitude across different symbolic notations (integers, decimals, and fractions). We hypothesized that, consistent with previous research, all of the number types would activate the IPS. The main questions concerned possible differences between the number types. If all number types activate the same abstract neural representation (based on relative rather than absolute magnitude, to take account of the scale difference between integers and rational numbers), then no differences among the number types would be expected. A second possibility is that neural activation of integers will differ from that of rational numbers

(either fractions or decimals), both because the latter are more complex and because the overall magnitude scale differs. A third possibility, based on the behavioral findings of DeWolf et al., 2014, is that fractions will evoke a neural signature distinct from that of either magnitude-equivalent decimals or integers, whereas the latter two number types will evoke similar activation patterns.

3.3 Methods

Participants Sixteen participants (12 female, mean age 21 years) with no documented history of neurological disorders were recruited at the University of California, Los Angeles (UCLA) through a flyer distributed in the Psychology department. Participants signed informed consent prior to the experimental session, and were paid $30 for their participation in the 1-hour study, in compliance with the procedures accepted by the local institutional review board (IRB).

Stimuli Stimuli consisted of pairs of numbers in one of three possible symbolic types: fractions (e.g., 1/2, 3/4), decimals (e.g., .50, .75) or integers (e.g., 50, 75). Within each pair, numbers were always of the same type. In order to control for the number of digits on the screen across symbolic types, only single-digit fractions, double-digit decimals, and double-digit whole numbers were presented. Thus, instances of the three symbolic number types were always constructed from exactly two digits. All of the fraction comparison pairs were comprised of fractions that did not have any common components. This constraint served to minimize the use of shortcut strategies, thereby encouraging participants to access the

holistic magnitude of each individual fraction. Magnitude-equivalent decimals were created by dividing out the corresponding fraction and rounding the result to two decimal places. Integers were created by multiplying the matched decimal by 100 to create a two-digit number. A total of 40 unique comparison pairs were generated for each number type. Because the numbers in the comparisons were shown sequentially, rather than simultaneously, each pair was shown twice, once in each order. Accordingly, there were a total of 80 trials for each of the three number types.

Behavioral Task Participants were given instructions before entering the scanning room, after performing a routine safety check. Participants were told that they would see a series of numbers presented sequentially in pairs. Each trial started with a fixation cross, at the middle of the screen, for .5 s followed by a brief blank screen jittered for .1-2 s. The first number was then presented for 1.5 s followed by a brief blank screen, which was jittered for 2-7 s, and then a second number (see Figure 3.1). Participants controlled the length of presentation of the second number by pressing a button to indicate whether the second number was larger or smaller than the first number. They were instructed to try to go as fast as possible without sacrificing accuracy. The 240 total trials (80 per symbolic type) were evenly distributed across four runs. Allocation of pairs across the four runs and order of presentation within each run was determined randomly for each participant.

fMRI Data Acquisition Data were acquired on a 3 Tesla Siemens Tim Trio Magnetic Resonance Imaging (MRI) scanner at the Staglin IMHRO Center for Cognitive Neuroscience at UCLA. Structural data were acquired using a T1-weighted sequence (MP RAGE, TR =

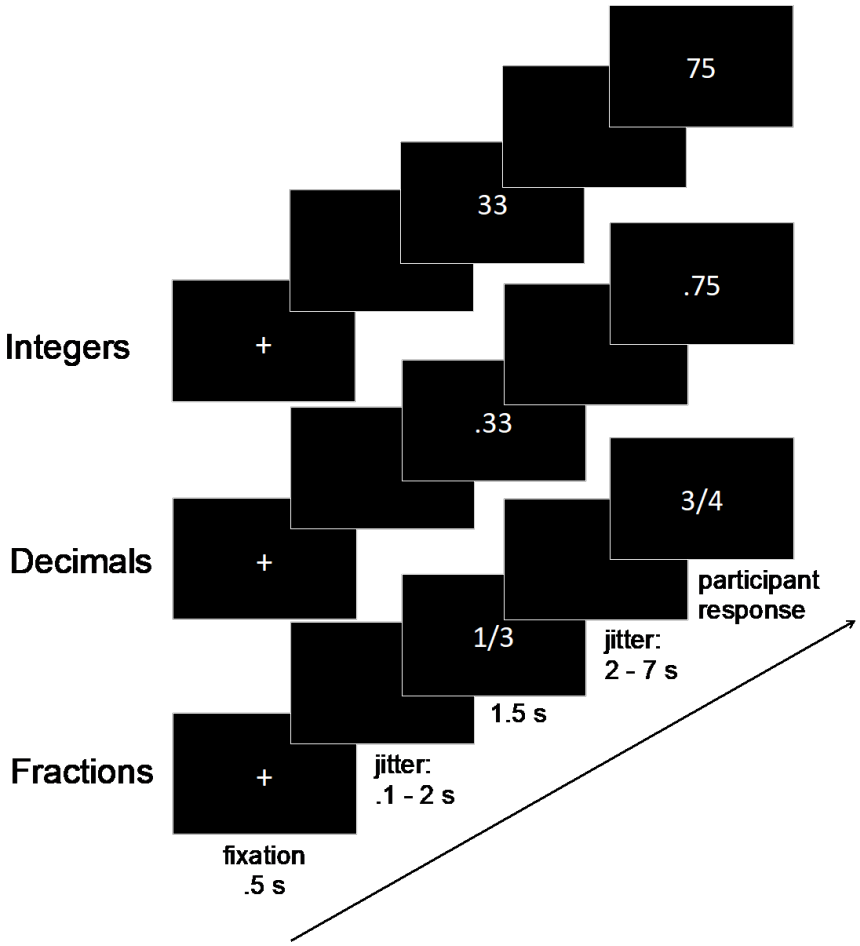

Figure 3.1: Trial procedure for each of the number type conditions (integers, decimals, fractions).

1,900 ms, TE = 2.26 ms, voxel size 1 mm3 isovoxel). Blood oxygenation level dependent (BOLD) functional data were acquired with a T2-weighted Gradient Recall Echo sequence (TR = 2,000 ms, TE = 30 ms, 32 interleaved slices, voxel size 3 × 3 × 4 mm, Flip Angle = 78 degrees). Overall, individual runs lasted an average of 566 s (min=492 s, max = 756 s).

Data preprocessing Data analysis was carried out using FSL (Smith et al., 2004). Prior to analysis, data underwent a series of conventional preprocessing steps including motion correction (Jenkinson et al., 2002), slice-timing correction (using Fourier-space time-series phase- shifting), spatial smoothing using a Gaussian kernel of 5 mm full-width half-max, and high-pass temporal filtering (Gaussian-weighted least-squares straight line fitting, with sigma = 50 s). Data from each individual run were analyzed employing a univariate general linear model approach (Monti, 2011) with pre-whitening correction for autocorrelation (Woolrich, Ripley, Brady, & Smith, 2001).

Univariate analysis For each run of each participant, a univariate GLM analysis was conducted with three regressors of interest marking the onset time and duration of the presentation of the first number of each pair, separately for each notation type (fractions, decimals, and integers). A number of additional regressors modeled the second number presentation, cue periods, and motion (first and second derivatives, and their difference). Data from the presentation of the second number was not analyzed further because it was confounded with movement (from pressing the response button) and cognitive processes relating to the comparison task (cf., Todd, Nystrom, & Cohen, 2013). For each run we computed seven contrasts. These were based on the data collected during the presentation

of the first number in a comparison pair. These included the simple effects of each notation type (fraction vs. baseline, decimal vs. baseline and integer vs. baseline), as well as the pairwise differences between them (fractions > decimals; fractions > integers; integers > decimals; and decimals > integers). Prior to group analysis, individual statistical maps were transformed into MNI template space via a 2-step procedure concatenating a boundary-based co-registration to align functional data to single-subject anatomical data and a 12 degrees of freedom linear co-registration to align single-subject anatomical data to the MNI template. Data from individual runs were aggregated using a mixed-effects model (i.e., employing both the within- and between-subject variance), using automatic outlier detection. Z (Gaussianised t) statistic images were thresholded using a cluster correction of $Z > 2.3$ and a (corrected) cluster significance threshold of $p = .05$.

In order to avoid reverse subtractions (Morcom & Fletcher, 2007), for each A>B contrast (e.g., fractions > decimals), we restricted the analysis to voxels for which the sum of the Z statistic associated with task A (compared to fixation; ZA) and the Z statistic associated with task B (compared to fixation; ZB) resulted in a number greater than zero (i.e., (ZA+ZB)>0). In other words, for a voxel to be included in the contrast analysis, either ZA and ZB had to both be positive values (in which case it is not possible to have reverse subtractions), or ZA had to be more positive than ZB was negative, thereby preventing the possibility of a brain activation resulting from a weak ZA coupled with a strongly negative ZB.

MVPA analysis The input to the multivariate pattern analysis (MVPA) was a set of volumes of regression coefficients (i.e., "beta" values) marking the magnitude of activation, for each voxel, in each trial (per participant). These trial-wise "patterns of activations" were

obtained by employing the iterative Least Squares – Separate approach (LS-S; Mumford et al., 2012) in which a separate GLM is run (here, using FILM with local autocorrelation; Woolrich et al., 2001) for each trial. The patterns of activation were then concatenated across time to construct a subject-wise "beta-series" of activation magnitude per trial per voxel (Rissman et al., 2004).

Representational Similarity Analysis (RSA) was run on the beta-series of activation magnitudes, in MATLAB using the RSA toolbox (Nili et al., 2014). RSA characterizes the representation in a brain region by a representational dissimilarity matrix (RDM), and compares the empirical matrix with a model. An RDM is a square symmetric matrix, with each entry referring to the dissimilarity between the activity patterns associated with two trials (e.g., entry (1,2) would represent the dissimilarity between activity patterns of trial 1 and trial 2 for a given participant). Each element of the RDM is calculated as 1 minus the Spearman correlation between the beta-series for each pair of trials. Models were manually generated to reflect idealized RDMs expected if the group of voxels was indeed modulating its activity with respect to the manipulation (see models in Figure 3.2). The Number Type Model (Figure 3.2a) was designed to test the overall ability to distinguish between each of the three number types. We then compared each of the pairwise number-type combinations to attempt to distinguish between each number type. The assumption behind the model RDMs was that a group of voxels sensitive to an experimental condition would display lower dissimilarity for same-condition trials as opposed to different-condition trials.

The RSA was performed with a searchlight approach (searchlight radius: 6mm or 2 voxels; cf. Kriegeskorte et al., 2006) within an anatomical mask of the IPS as defined by the Jülich

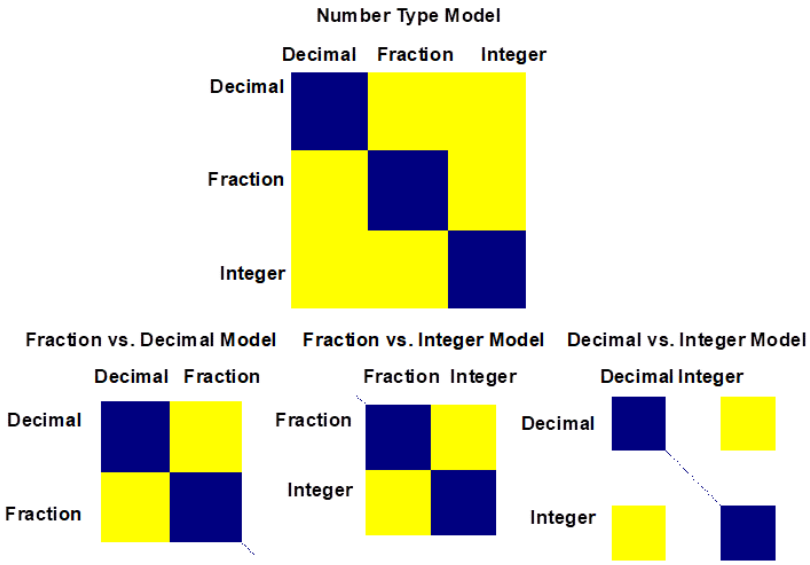

Figure 3.2: Ideal models generated for the RSA searchlight MVPA. Each matrix represents a dissimilarity matrix where yellow (1) denotes completely dissimilar items and blue (0) denotes maximally similar items.

Histological Atlas (available in FSL; Choi et al., 2006; Scheperjans et al., 2008). Within each searchlight sphere, a Spearman coefficient was computed between the empirical and model RDMs, yielding a single second-order similarity value per voxel, which reflected the resemblance of searchlight sphere activity with the hypothesized model. These coefficients were registered to the standard template, with the same 2-step procedure employed for univariate single-subject statistical parametric maps, and assessed for significance ($\rho > 0$) using FSL's randomize with threshold-free cluster enhancement (corrected $p < .05$) (Smith & Nichols, 2009; Winkler et al., 2014).

3.4 Results

3.4.1 Behavioral Results

Mean accuracy on the magnitude comparison task for each number type was obtained by averaging over all participants. A one-way repeated measures ANOVA revealed a significant effect of number type ($F(2, 30) = 23.23$, $MSE = .002$, $p < .001$), with fractions having lower accuracy than decimals (fractions: 84% vs. decimals: 92%, $t(15) = 6.72$, $p < .001$) and integers (91%, $t(15) = 4.82$, $p < .001$). There was no difference in accuracy between decimals and integers ($t(15) = .69$, $p = .50$).

Mean response times (RTs) for correct trials were averaged for each number type across participants. A one-way repeated measures ANOVA revealed a significant effect of number type ($F(2, 30) = 24.34$, $MSE = .09$, $p < .001$), with fractions being compared more slowly than either decimals (fractions: 1.91 s vs. decimals: 1.30 s, $t(15) = 5.22$, $p < .001$) or integers (1.24 s, $t(15) = 5.19$, $p < .001$). There was no significant difference in response time between decimals and integers ($t(15) = 1.15$, $p = .27$). Because accuracy was at ceiling for both decimals and integers, we focused on distance effects based on RTs for correct responses. Figure 3.3 shows the average RT for each trial across participants for fractions, decimals, and integers. In order to assess the distance effect, we conducted regression analyses for response times based on a logarithmic distance measure (i.e., log(|first number – second number|), which we will abbreviate as "log Dist" (see DeWolf et al., 2014; Hinrichs et al., 1981). Log Dist significantly predicted RT outcomes for each of the number types (fractions: β = -64, $t(37) = 5.01$, $p < .001$; decimals: β = -33, $t(37) = 2.16$, $p = .04$; integers: β = -33, $t(37) =$

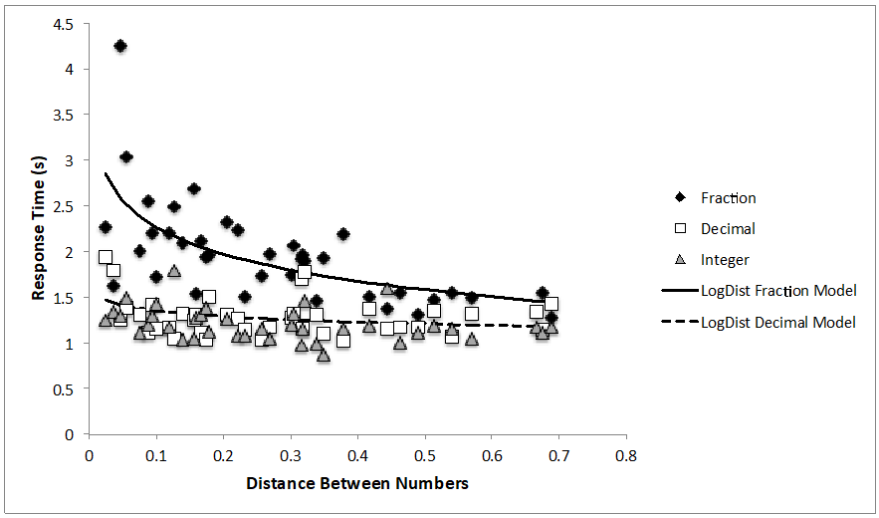

Figure 3.3: Average correct response times for each trial across participants for fractions, decimals, and integers. Fitted lines represent predictions derived from LogDist models for Fractions and Decimals. (Because predictions of the Integer model were nearly identical to those of the Decimal model, the Integer model is excluded here for simplicity.)

2.12, p = .04). These results replicate the pattern of distance effects observed by DeWolf et al. (2014), including (as evidenced by the much larger beta coefficient for fractions) a more pronounced distance effect for fractions than for either of the other two number types.

3.4.2 fMRI Results

Univariate analyses The contrast of fractions versus decimals resulted in extensive activations within and around the left horizontal segment of the intraparietal sulcus, spanning inferior (Brodmann Area [BA] 40) and superior (BA 7) parietal lobuli, as well as the junction of the intraparietal and intraoccipital sulci. Additional left hemispheric activations were detected in frontal cortex, centered around the precentral gyrus (BA 6) together with smaller foci within the superior (BA 6) and middle (BA 9) frontal gyri, and in temporal cortex,

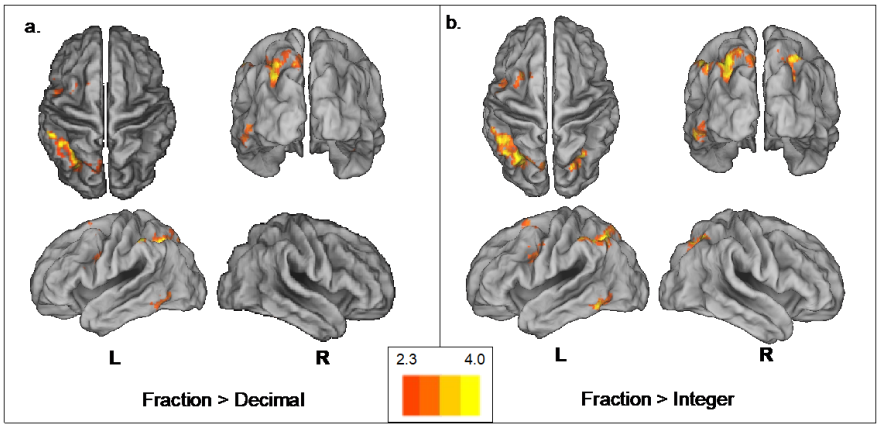

Figure 3.4: Results of the univariate analysis for (a) comparison of fraction and decimal activation and (b) comparison of fraction and integer activation, from dorsal, posterior, and lateral views. Red areas represent significant differences in activations. The color scale represents z-values for significant activations.

spanning the most caudal segments of the inferior and middle temporal gyri (BA 37). Finally, right lateralized activations were observed in the cerebellum, with foci in Crus I and Lobules VI and VIIB (see Figure 3.4a and Table 3.1 for complete list of local maxima).

MNI Coordinates					
x	y	z	Hem	Anatomical Label (BA)	Z
Parietal					
-42	-48	48	L	Inferior Parietal Lobule (hIPS; 40)	4.30
-30	-58	44	L	Superior Parietal Lobule (hIPS; 7)	3.93
-54	-38	48	L	Inferior Parietal Lobule (40)	3.91
-28	-70	38	L	Occipito-Parietal Junction (hIPS/IOS;40/7)	3.75
-28	-76	54	L	Superior Parietal Lobule (hIPS; 7)	3.71

-34	-50	42	L	Inferior Parietal Lobule (hIPS; 40)	3.71
Frontal					
-56	14	24	L	Precentral Gyrus/Inferior Frontal Gyrus (6/44)	3.77
-50	6	32	L	Precentral Gyrus (6)	3.65
-22	6	70	L	Superior Frontal Gyrus (6)	3.27
-30	2	62	L	Precentral Gyrus/Superior Frontal Gyrus (6)	3.25
-34	2	28	L	Precentral Gyrus/Inferior Frontal Gyrus (6/44)	3.18
-36	2	62	L	Precentral Gyrus/Superior Frontal Gyrus (6)	3.16
-54	8	46	L	Precentral Gyrus (6)	2.93
-40	-4	56	L	Precentral Gyrus (6)	2.93
-28	-8	58	L	Precentral Gyrus (6)	2.91
-18	12	66	L	Superior Frontal Gyrus (6)	2.85
-50	26	28	L	Inferior Frontal Gyrus/Middle Frontal Gyrus (44/9)	2.70
-44	-4	30	L	Precentral Gyrus (6)	2.68
Temporal					
-50	-56	-22	L	Inferior Temporal Gyrus (37)	3.18
-56	-56	-14	L	Inferior Temporal Gyrus (37)	3.13
-48	-64	-4	L	Inferior Temporal Gyrus (37)	3.03
-50	-60	0	L	Middle Temporal Gyrus (37)	2.84
-48	-56	0	L	Middle Temporal Gyrus (37)	2.60
Subcortical					
38	-56	-32	R	Crus I	3.88

38	-60	-40	R	Crus I	3.41
28	-76	-50	R	Lobule VIIB	3.19
32	-72	-26	R	Crus I	3.13
44	-66	-30	R	Crus I	2.86
26	-60	-34	R	Lobule VI	2.86

Table 3.1: Local Maxima for the Fractions > Decimals Univariate Contrast (Abrev.: hIPS: horizontal segment of the intraparietal sulcus; IOS: intraoccipital sulcus; L: left; R: Right)

The contrast of fractions versus integers resulted in extensive activations in bilateral parietal cortex (with L>R), centered within and around the horizontal segment of the intraparietal sulci, spanning inferior (BA 40) and superior (BA 7) parietal lobuli.1 Similarly to the contrast of fractions versus decimals, left lateralized activations were also obtained in frontal cortices, mostly within the precentral gyrus (BA 6) together with foci across superior (BA 6, 8) and middle (BA 8) frontal gyri, and in the caudal section of temporal cortex, in the inferior and middle temporal gyri (BA 37). Finally, right hemispheric activations were again observed in the cerebellum, with foci in Crus I and II, and Lobule VI (see Figure 3.4b and Table 3.2 for a complete list of local maxima).

Direct comparison of the decimal and integer conditions, in both directions (i.e., decimals > integers; integers > decimals), failed to reveal any significant activation.

MNI Coordinates

x	y	z	Hem	Anatomical Label (BA)	Z
Parietal					
-28	-62	46	L	Superior Parietal Lobule (hIPS; 7)	4.71
-30	-74	54	L	Superior Parietal Lobule (hIPS; 7)	4.39
-30	-70	54	L	Superior Parietal Lobule (7)	4.21
-48	-50	52	L	Inferior Parietal Lobule (40)	3.99
-48	-48	48	L	Inferior Parietal Lobule (40)	3.98
-40	-46	46	L	Inferior Parietal Lobule (hIPS; 40)	3.82
32	-68	48	R	Inferior Parietal Lobule (40)	3.82
30	-60	50	R	Superior Parietal Lobule (hIPS; 7)	3.74
28	-64	50	R	Inferior Parital Lobule (40)	3.69
24	-68	48	R	Superior Parietal Lobule (7)	3.58
42	-54	58	R	Superior Parietal Lobule (7)	3.40
22	-72	50	R	Superior Parietal Lobule (7)	3.33
Frontal					
-50	10	28	L	Precentral Gyrus (6)	3.88
-52	12	24	L	Precentral Gyrus (6)	3.78
-52	10	36	L	Precentral Gyrus (6)	3.62
-24	16	54	L	Superior Frontal Gyrus (8)	3.62
-36	4	28	L	Precentral Gyrus (6)	3.51
-24	12	58	L	Middle Frontal Gyrus (8)	3.51
-46	6	34	L	Precentral Gyrus (6)	3.41

-36	2	62	L	Precentral Gyrus (6)	3.37
-52	2	42	L	Precentral Gyrus (6)	3.34
-18	12	64	L	Superior Frontal Gyrus (6)	2.97
-24	8	46	L	Middle Frontal Gyrus (8)	2.82
-24	14	66	L	Superior Frontal Gyrus (8)	2.78
Temporal					
-50	-56	-10	L	Inferior Temporal Gyrus (37)	3.76
-54	-54	-14	L	Inferior Temporal Gyrus (37)	3.65
-42	-60	-6	L	Inferior Temporal Gyrus (37)	3.08
-56	-64	-10	L	Inferior Temporal Gyrus (37)	2.98
-48	-64	-4	L	Inferior Temporal Gyrus (37)	2.96
-48	-48	-14	L	Inferior Temporal Gyrus (20/37)	2.76
Subcortical					
38	-56	-30	R	Crus I	3.87
36	-64	-44	R	Crus II	3.25
40	-60	-38	R	Crus I	3.24
20	-66	-26	R	Lobule VI	3.21
38	-58	-42	R	Crus II	3.18
32	-72	-26	R	Crus I	3.18

Table 3.2: Local Maxima for the Fractions > Integers Univariate Contrast (Abrev.: hIPS: horizontal segment of the intraparietal sulcus; L: left; R: Right)

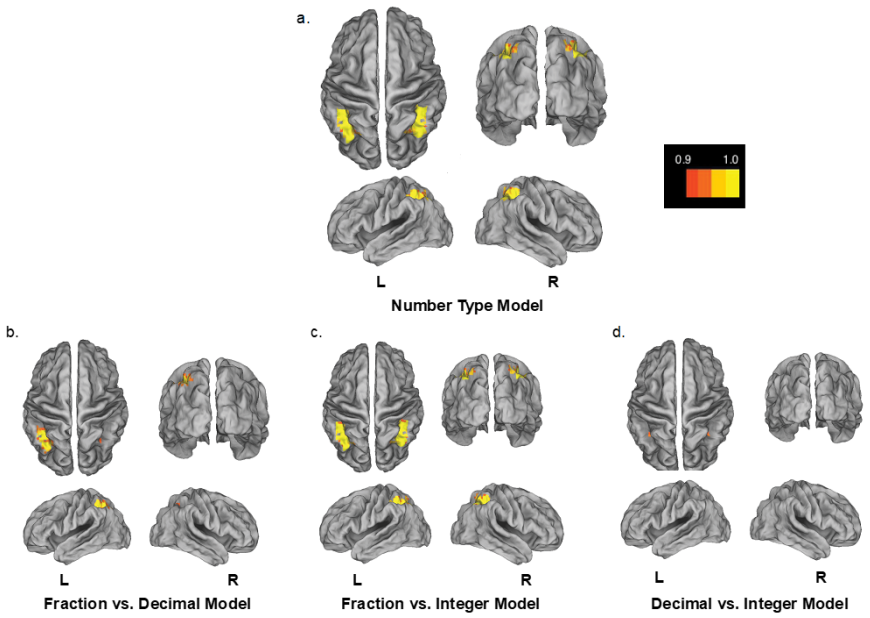

Figure 3.5: Results of the multivoxel pattern analysis (MVPA) from dorsal, posterior, and lateral views for each of the four hypothesized models (see Figure 2). The color scale represents 1 - p-values (e.g., .95 to 1 would be significant). Note: The searchlight analysis was restricted to the IPS, which was selected as a region of interest.

Multivariate analyses Because MVPA requires an equal number of trials across all conditions, one participant was excluded from this analysis because she did not finish one of the runs due to a computer error (missed two trials).

Figure 3.5 shows the areas within the IPS that yielded significant activations for each of the four models. The Number Type model (distinguishing between the three number types) shows a broad set of bilateral activations. Mirroring the results of the univariate analysis, the Fraction vs. Decimal model shows mostly left-lateralized IPS activation, whereas the Fraction vs. Integer model shows bilateral IPS activation. Unlike the results of the univariate

analysis, the Decimal vs. Integer model yielded a small number (3) of significant voxels that distinguished between decimals and integers. While this is a small area, it points to a possible pattern difference in the encoding of decimals and integers beyond what the univariate analysis revealed.

3.5 Discussion

Fraction Magnitudes are Neurally Distinct from Decimals and Integers The central goal of the present study was to distinguish between possible models of neural representation for different symbolic number formats. The behavioral results showed that each of the number types elicited a reliable distance effect on correct RT. The presence of a distance effect suggests that all number types were processed holistically. However, the neuroimaging results showed that magnitudes evoked distinct neural patterns that distinguished the number types. Results of both a univariate analysis and MVPA indicate that while fractions, decimals, and integers all activate areas of the IPS, fractions yield a distinct pattern of activation associated with a unique subarea of the IPS. In contrast, decimals and integers yielded very similar and overlapping patterns, with MVPA identifying only a very small set of voxels that distinguished the latter two number types. These results suggest that while neural representations, across notations, all elicit activation within the intraparietal sulcus, neural representation appear to be sensitive to number representation (notably, base-10 numbers versus fractions), but not to number type (natural versus rational).

To our knowledge, the present neuroimaging study is the first to compare fractions with both decimals and integers. The two previous studies (Ischebeck et al., 2009; Jacob & Nieder,

2009a) that investigated the representation of fraction magnitudes using fMRI had assumed that because fractions activate the IPS (as do integers), and because fraction activation was modulated by a distance effect based on holistic magnitude, the brain represents proportional (fraction) magnitudes in the same way that it does absolute (integer) magnitudes. However, by making direct comparisons among all three number types, the present study was able to clearly dissociate magnitude activations for fractions as compared to those for either integers or decimals.

Isolating Magnitude Representations for Individual Numbers An important methodological innovation of the present study is its use of a design based on sequential presentation of individual numbers in a magnitude comparison task. Compared to passive observation of numbers, the magnitude comparison task strongly guides participants to access holistic magnitude representations for individual numbers. Moreover, the behavioral results from the comparison task fully replicated previous work comparing performance with the three number types (DeWolf et al., 2014). Comparisons were less accurate and slower for pairs of fractions than for pairs of decimals or integers. A distance effect was obtained for all number types, but was most pronounced for fractions. Our behavioral results thus confirm that participants in our neuroimaging paradigm were performing magnitude comparisons in essentially the same way as has been observed in previous behavioral studies.

At the same time, the sequential nature of the present design allowed us to decouple the process of accessing a magnitude representation for an individual number from the process of magnitude comparison. Our fMRI analyses focused solely on the initial 1.5s period when a single number was displayed. During this period participants were motivated to access

the magnitude of the presented number, but were unable to initiate a comparison because the second number in the pair had not yet appeared. Previous neuroimaging studies with fractions recorded neural signals during the comparison process itself. In contrast, our findings provide a clear picture of the neural activity underlying access to the magnitude of a single individual number, isolated from the additional activity that would be triggered by comparing two magnitudes.

What is Special About Fraction Magnitudes? We considered three hypotheses about the relation between magnitude representations for different symbolic notations. (1) All notations might evoke some universal, fully abstract magnitude code; (2) the magnitude code might differ between natural numbers (integers) and the more complex rational numbers (fractions and decimals); or (3) the magnitude code for fractions might differ from that for the base-10 notations (decimals and integers). Our findings clearly support the third of these hypotheses. To the best of our knowledge, no previous study has shown such a strong dissociation between the neural patterns elicited by alternative notations for the same magnitude. Even though 2/5 and .40 express the same magnitude, the brain processes the two symbols very differently. In contrast, the magnitude representations for a decimal (.40) and an integer expressing a magnitude 100 times larger (40) are very similar. Importantly, the latter result implies that the neural code for numerical magnitude is on a scale that is fundamentally relative rather than absolute. Thus base-10 notations evoke similar activation patterns based on their relative magnitudes, whereas the bipartite fraction notation is processed very differently from either.

Future Directions The present study lays the groundwork for further exploration of the differences among neural representations evoked by different symbolic number types. Behavioral evidence points to a major conceptual distinction between fractions and decimals, with the former being selectively used to code the magnitudes of discrete entities (which can be counted), and the latter selectively used to code the magnitudes of continuous quantities (which can be either estimated or measured by imposing arbitrary units; see Rapp, Bassok, DeWolf, & Holyoak, 2015). In addition, it is important to examine neural processing in mathematical tasks other than those that focus on magnitudes. Whereas fractions are disadvantaged relative to decimals in magnitude comparison tasks, fractions convey reliable advantages in a variety of reasoning tasks. Because of their bipartite structure, fractions have a much more natural correspondence to relational concepts based on ratios of countable sets (DeWolf, Bassok, & Holyoak, 2015a). The relational aspects of fraction representations appear to make fraction understanding a critical bridge to learning algebra (DeWolf, Bassok, & Holyoak, 2015b), which depends critically on grasping the concept of a variable (understood to represent a quantity of unknown magnitude). The "isolation" technique introduced in the present paper (imaging activity evoked by an individual number as the participant prepares for a specific mathematical task performed immediately afterwards) might usefully be extended to compare the neural patterns evoked by the same symbol (e.g., a fraction) in preparation for tasks that require different types of information (e.g., magnitudes or relational concepts).

Footnote 1 The analyses comparing fractions to decimals and fractions to integers revealed hemispheric differences, with the latter comparison resulting in bilateral parietal activations

and the former resulting in left lateralized activations only. However, these differences were mainly attributable to the non-linear nature of the thresholding procedure. Inspection of uncorrected statistical parametric maps resulting from the fractions-minus-decimals contrast revealed clusters of above-threshold voxels (i.e., individual $Z > 2.3$); however, these were too small to survive the cluster-extent thresholding. These subthreshold activations explain why no difference was apparent when directly comparing decimals and integers.

Acknowledgements Preparation of this paper was supported by pilot funding from the Staglin IMHRO Center for Cognitive Neuroscience at UCLA, and by NSF Fellowship DGE-1144087 to MD. We thank Michael Ambrosi and Jingqi Yu for assistance with collecting fMRI data.

CHAPTER 4

Neurocomputational models predict brain activity during analogical reasoning

4.1 Introduction

A key component of human reasoning, creativity, and problem solving is the ability to draw inferences based not only on individual concepts, but also on relations between concepts. A paradigmatic example of relational processing is analogical reasoning (Holyoak, 2012). To determine whether or not a verbal analogy in the canonical form *A:B :: C:D* (e.g., *rob:steal :: cry:weep*) is valid, two basic computations are required. First, it is necessary to represent the relation or relations between A and B (e.g., *synonym*), and that between C and D. Second, it is necessary to compare the two relations to determine whether they are the same (a valid analogy) or different (invalid).

Previous work on the neural substrate of analogical reasoning has focused on the process of relational comparison, which has been localized to a primarily fronto-parietal network (Vendetti & Bunge, 2014; Wendelken, Ferrer, Whitaker, & Bunge, 2016). Notably, the left rostrolateral prefrontal cortex (rlPFC), corresponding to Brodmann areas 10 and 47, is activated in a graded manner in response to increases in the complexity of the relations

being compared (Christoff et al., 2001; Bunge, Helskog, & Neuroimage, 2009) or the semantic distance between the A:B and C:D concepts (Green, Kraemer, Fugelsang, Gray, & Dunbar, 2010, 2012; for meta-analysis see Vartanian, 2012; Hobeika, Capucine, Garcin, Levy, & Volle, 2016). For verbal analogy tasks, additional frontal regions (Brodmann areas 44, 45, 46, and 6) appear to be involved in retrieving relational and abstract information (Bunge, Wendelken, Badre, & Wagner, 2005; Della Rosa, Catricalà, Canini, Vigliocco, & Cappa, 2018). Parietal activations, particularly around the intraparietal sulcus and supramarginal gyrus, have been associated with relational reasoning across several domains (see Wendelken, 2015).

The neural substrate of relation representations has received much less attention. Although progress has been made in mapping semantic representations of individual words and word combinations (Pereira et al., 2018), major hurdles have impeded progress in determining the neural basis for the kinds of abstract and structured relations that underlie analogical reasoning. In particular, it is difficult to distinguish the representation of a relation (e.g., *opposite*) from that of the individual concepts being related (e.g., *friend:enemy*), particularly because verbal analogies do not provide a relation name as part of the linguistic input.

Recent advances in machine learning provide potential theoretical tools for investigating the neural basis of relation representations. In the present paper, we applied neurocomputational models to guide an investigation of the neural substrate for analogical reasoning. The three models considered here (see Methods) are based directly or indirectly on semantic representations of individual words created using a deep-learning algorithm, Word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Le & Mikolov, 2014). After learning to

predict words in local contexts taken from a large text corpus, Word2vec yields a vector of semantic features (300 features, each with continuous values) for individual words. Recent work has shown that these representations can predict the similarity of neural responses to individual words, which we will term first-order similarity (Pereira et al., 2018). However, additional computations are necessary to model relation representations, which may involve second-order similarity. Consider a word pair such as *friend:enemy*. A first-order representation of the meaning of the pair can be formed by simply concatenating the feature vectors for the two individual words (creating a vector of length 600). We term this model, directly calculated from Word2vec vectors, Word2vec-concat.

To represent the general relation between any two words, Mikolov et al., 2013 suggested using the difference between the vectors of the individual words (a difference vector of length 300). This model, termed Word2vec-diff, treats the difference vector as a general measure of the second-order semantic relation between a pair of words (e.g., the same basic measure is used to represent both friend-enemy and rob-steal, even though the specific relations differ in the two cases).

A third model, BART (*Bayesian Analogy with Relational Transformations*; Lu, Chen, & Holyoak, 2012), aims to represent the specific relation between each pair of words. BART takes as inputs pairs of Word2vec vectors labeled as positive or negative examples of a specific relation, and applies Bayesian regression to derive a weight distribution predicting the posterior probability that any word pair exemplifies that relation. The model is trained on 79 abstract semantic relations (from a linguistic taxonomy created by Bejar, Chaffin, & Embretson, 1991), using example pairs taken from norms collected by Jurgens, Turney,

Mohammad, & Holyoak, 2012. After learning, BART represents the specific relation between any two words as a vector of posterior probabilities across the 79 relations. The model captures the intuition that many word pairs involve multiple relations to some degree (e.g., friend-enemy mainly exemplifies the relation opposite, but also involves a type of similarity in that both words refer to social categories). The three models thus allow us to test three alternative hypotheses regarding human relation representations: patterns of neural responses may be based solely on first-order similarity (Word2vec-concat); or on a general measure of second-order similarity (Word2vec-diff); or on a measure of specific relation(s) that contribute to second-order similarity (BART).

In the present study, we used univariate and multivariate methods to investigate the underlying neural activity during two critical stages of analogical reasoning: relation representation and relational comparison. We employed a sequential design (see DeWolf, Chiang, Bassok, Holyoak, & Monti, 2016 to separate the two operations. In a rapid event-related fMRI design, 16 healthy volunteers were asked to evaluate two pairs of semantic concepts. Each analogy was presented as two pairs of words, an *A:B* pair (e.g., up:down) followed by a *C:D* pair (e.g., *high:low*) exemplifying the same relation as the *A:B* pair (valid analogy), or else a *C':D'* pair (e.g., *couch:sofa*) exemplifying a different relation (invalid). All analogies were based on word pairs exemplifying three abstract relation types (*similar, contrast, and cause-purpose*) from the Jurgens et al., 2012 norms (see Table 1). The *A:B* phase provided a relatively pure measure of neural activity involved in coding the *A:B* relation, whereas the *C:D* phase included neural activity required to maintain the *A:B* relation, form the *C:D* relation, and compare the two.

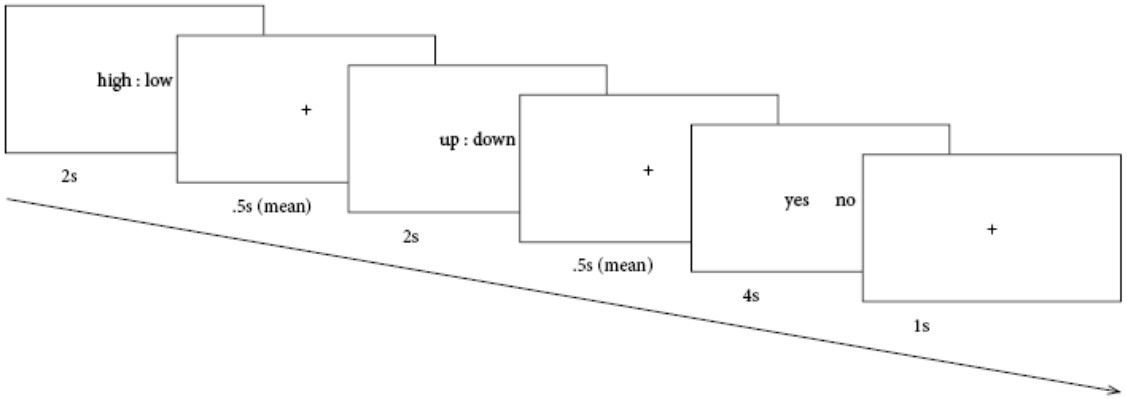

Figure 4.1: Subjects were shown two word pairs, first an *A:B* pair for 2 seconds, then a *C:D* pair for 2 seconds after a jitter. Subjects responded by pressing a button box, but the location of "yes" and "no" varied from trial to trial.

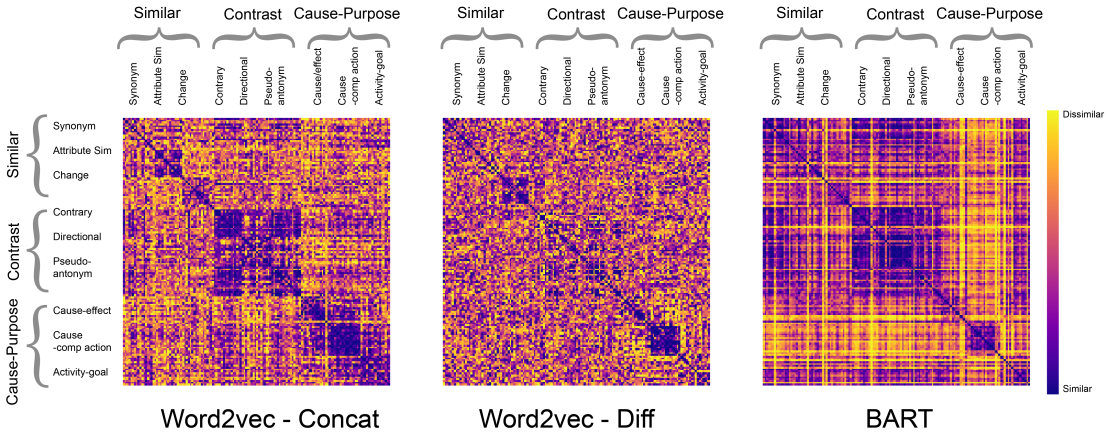

Figure 4.2: The first three models from left to right were derived from the computational models. The categorical model (right) reflects the Jurgens et al. (2012) norms.

4.2 Results

Localization of relational reasoning We computed the main effects of the *A:B* and *C:D* phases in order to uncover regions engaged in relation representation (Figure 4.3). The main effects of the *A:B* and *C:D* stage of each trial were associated with mostly left-lateralized frontal and temporal activity, bilateral parietal activity, and activity in the occipital lobe. *A:B* trials were uniquely associated with clusters of increased activation in the supplementary motor cortex and caudate (for a full list consult table B.1), while *C:D* trials were uniquely associated with clusters of increased activity in left inferior frontal gyrus as well as bilateral superior parietal cortex (for a full list consult table B.2). *A:B* and *C:D* analogical stimuli shared coactivations in the inferior lateral occipital cortex, fusiform gyrus and left frontal regions spanning the rostrolateral prefrontal cortex (Brodmann areas 10 and 47).

A direct univariate subtraction was run between the *C:D* and *A:B* phases to identify regions engaging in relational comparison (Figure 4.4). Notable clusters were observed in the left inferior frontal gyrus, bilateral superior parietal regions, fusiform gyrus, and rostrolateral prefrontal cortex. Local maxima are listed in supplementary table B.3.

Neural distribution of abstract relation representation We applied a searchlight (Kriegeskorte et al., 2006) classification analysis to determine where information relevant to abstract relation representation could be decoded. As shown in Figure 4.5, the main relationships (synonym, antonym, and cause-effect) could be distinguished during the *A:B* trials in left frontal, left temporal, and bilateral parietal cortices. The relationships could be distinguished during the *C:D* trials in many of the same regions, with additional discrim-

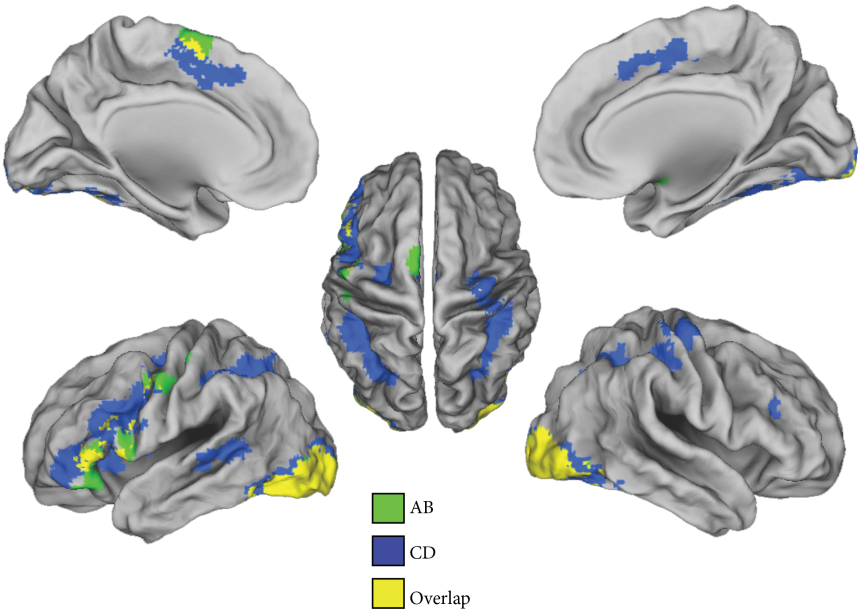

Figure 4.3: Main effects during *A:B* and *C:D* phases

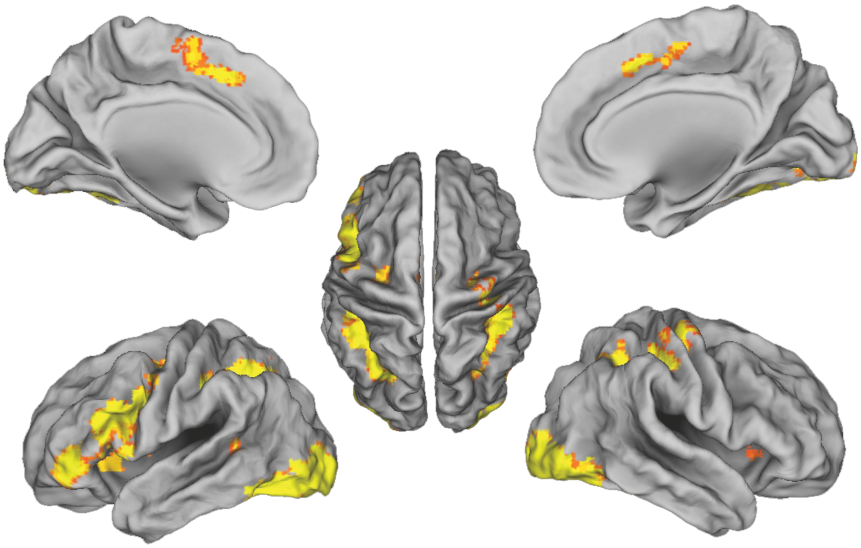

Figure 4.4: CD - AB.

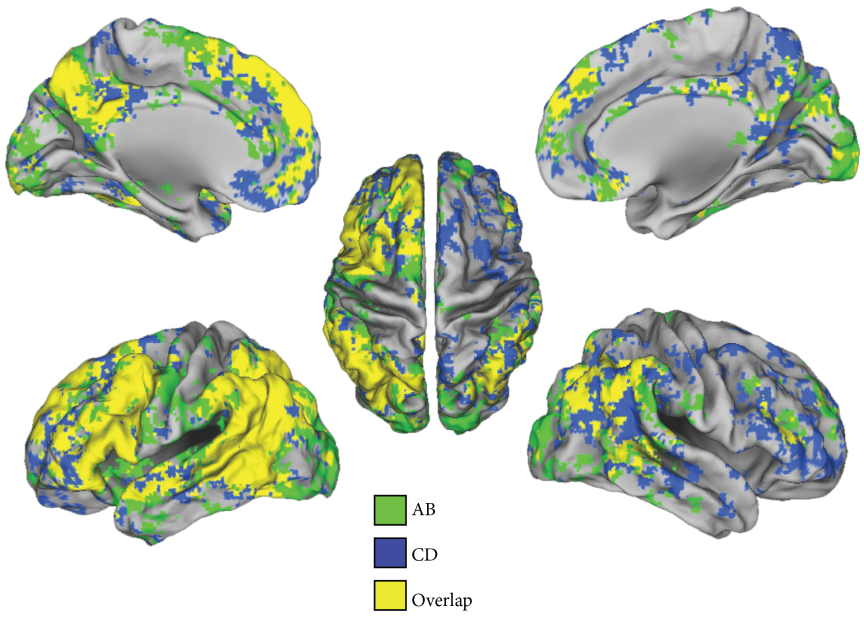

Figure 4.5: Local multivariate results. Significant regions indicate relational representation inability in the right hemisphere. These areas broadly overlapped across the two stages in the left prefrontal, left temporal, and bilateral parietal cortices.

We ran an ROI-based classification analysis to identify regions engaged in relational comparison during the *C:D* phase. Classifiers were trained to discriminate between *C:D* (valid) and C':D' (invalid) trials. As shown in Figure 4.6, frontal regions (BAs 10, 44, 45, 46, 47), the inferior parietal regions (aSMG, pSMG, AG), and the left posterior MTG displayed significantly above chance performance ($p < 0.05$).

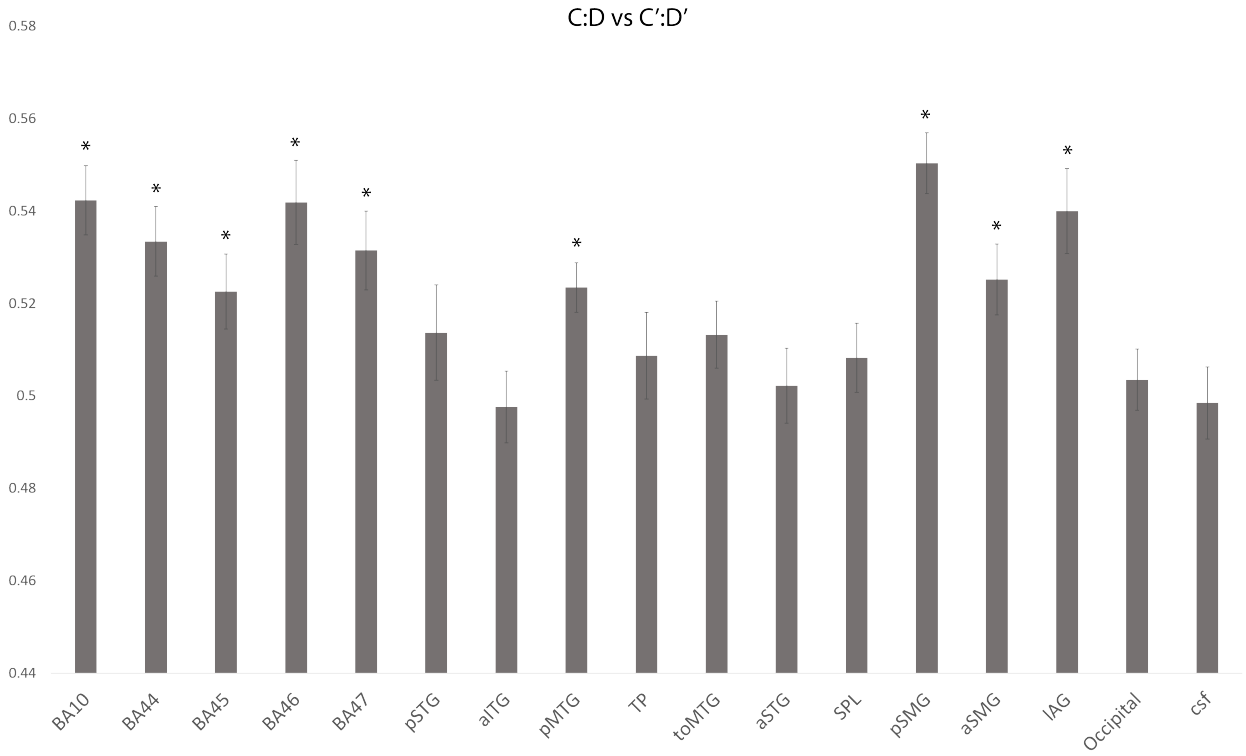

Figure 4.6: ROI ability to discriminate between valid and invalid trials

Computational accounts of relational reasoning Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008) was used to characterize the similarities of neural responses across word pairs. RSA characterizes the representation in a region of interest (ROI) by a trial-by-trial representational dissimilarity matrix (RDM), and compares this empirical matrix with a theoretical model (Carota, Kriegeskorte, Nili, & Pulvermüller, 2017; Nili et al., 2014).

For each ROI (see Methods), empirical RDMs were compared with idealized models shown in Figure 4.2. RDMs derived from the left IFG (BAs 44 and 45), the parietal cortex (pSMG), and the temporo-parietal junction (toMTG) were significantly correlated with all three models. Other parietal ROIs (SPL and aSMG) were significantly correlated with the Word2vec difference model and Bayesian model representations.

RDMs from each ROI were also compared with a categorical hypothesis model, created using the labels provided by the Jurgens et al., 2012 norms. This model was correlated with activity in frontal (BAs 44, 45, 47), posterior temporal (pMTG, pSTG, toMTG), and parietal regions (lSPL, lAG, lSMG, pSMG).

Within the left SPL, the BART-derived RDM was significantly more correlated with the neural RDM than both the Word2vec difference RDM and the categorical RDM (Figure 4.2).

For each analogy, that is, every valid *A:B::C:D* trial, a *relational similarity* measure was calculated by taking the cosine distance between word pair representations. Relational similarity in the context of these analogies is related to difficulty- the farther apart *A:B* and *C:D* are in relation representational space, the harder the analogy ought to be. Relational

similarity scores for each trial were correlated with voxel activity to identify brain regions that track this measure.

Relational similarity was computed, using the three different model representations, for each trial and was correlated (using Spearman's rho) with the corresponding beta-values during the *C:D* (valid) phase to produce a correlation map. Significance was assessed using FSL randomise (Winkler et al., 2014) with TFCE cluster correction (Smith & Nichols, 2009).

Regions correlated with the BART-derived relational similarity values are shown in Figure 4.8. BART predictions were significantly negatively correlated with voxel activity primarily in the left frontal cortex, left supramarginal gyrus and left posterior middle temporal gyrus. No regions were significantly correlated (after threshold-free cluster correction) with the two Word2vec derived relation representations.

4.3 Discussion

By separating the analogy task into two temporally distinct phases, our sequential design allowed us to separately study the neural operations underlying relation representation (e.g., the retrieval or encoding of the abstract relation itself) and relational comparison (e.g., evaluating whether *A:B* and *C:D* are of the same relation - DeWolf et al., 2016). We discuss our results in the context of these two phases below, and then relate them to our computational analysis to provide an account for the operations being carried out by previously reported regions.

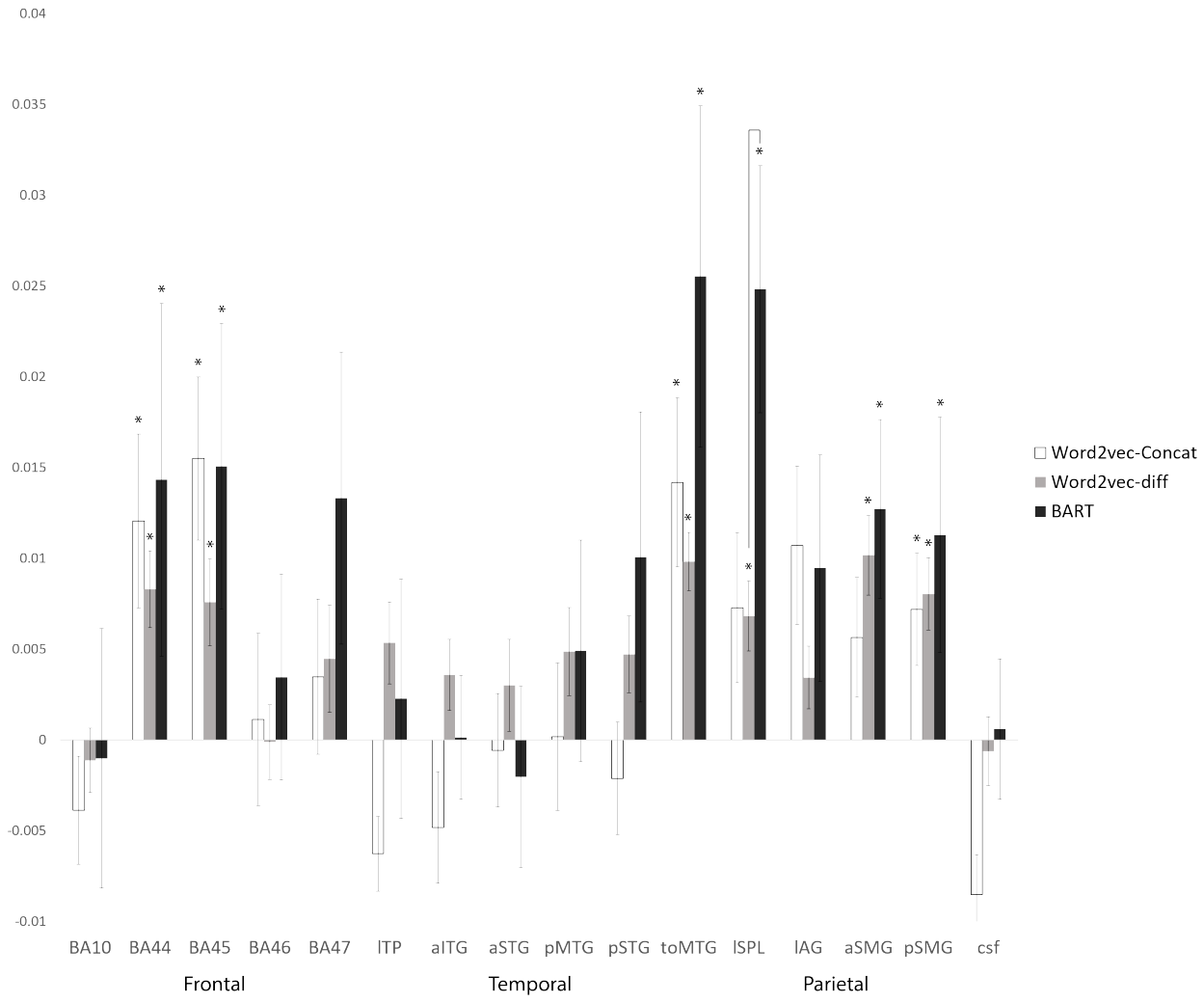

Figure 4.7: Each bar represents the average Spearman correlation between the empirical RDM and hypothesis model

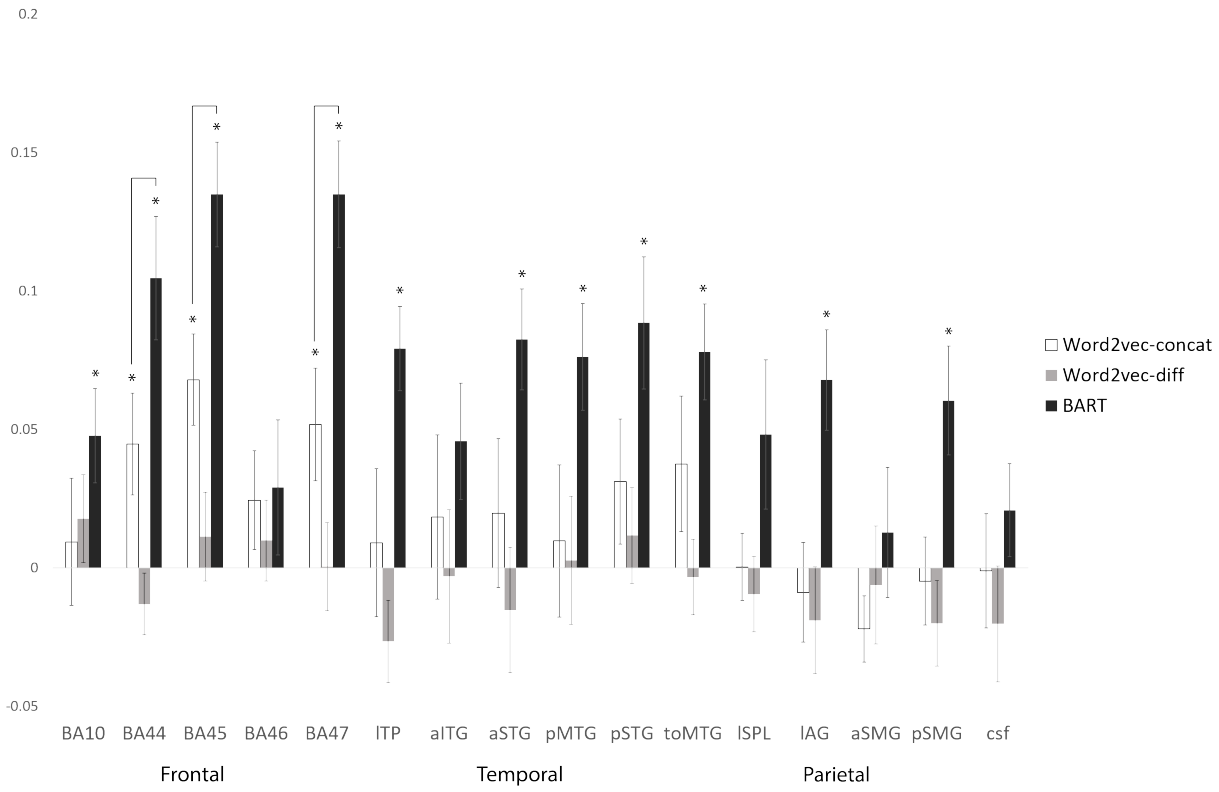

Figure 4.8: ROI analysis correlating mean voxel activity (within the ROI) with relational similarity calculated from the three models. Word2vec-diff did not explain ROI activity in any regions, while Word2vec-concat was correlated with mean ROI activity in frontal ROIs. BART was significantly correlated with ROI activity in frontal, parietal, and temporal regions, and significantly more correlated than Word2vec-concat.

Neural distribution of abstract relations One primary aim of this study was to localize and characterize relation representation in the brain. By analyzing the fMRI activity measured when reading the first word pair, we captured cognitive processes associated only with representing the first-order relation between words A and B, as described in DeWolf et al., 2016.

The searchlight analysis shows that the abstract relations used in this study are represented in a primarily left-lateralized fronto-temporo-parietal network. This suggests that higher-order relations (which may not arise intuitively from semantic features) are encoded to some degree within the same "semantic network" found in studies of single-word semantics (Binder, Desai, Graves, & Conant, 2009; Carota et al., 2017; de Heer, Huth, Griffiths, Gallant, & Theunissen, 2017; Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016). However, whether the widespread distribution of these representations are due to truly abstract relational representations or simply due to the linguistic nature of our task is a subject for future study.

Neural correlates of relational comparison The sequential design of this study was also tailored to investigate the neural correlates of relational comparison. We analyzed the difference between processing the C:D word pair and the A:B word pair using a univariate contrast analysis, and probed whether more active regions encoded information meaningful to comparing the two relations using a multivariate classification, following up with a voxelwise correlation analysis.

Within the prefrontal cortex, the univariate subtraction in Figure 4.4 revealed significant clusters in the left rostrolateral PFC, replicating prior results showing that the rlPFC is more

active during higher-order relational comparison (Christoff et al., 2001; Bunge et al., 2009). Consistent with this account, the multivariate patterns within the frontal pole (BA10) and left rlPFC (BA47) were able to discriminate between the valid and invalid trials, further highlighting their involvement in analogy.

Within the prefrontal cortex, the univariate subtraction in Figure 4.4 revealed significant clusters in the left frontal pole and left inferior frontal gyrus, corroborating previous results that the left rostrolateral PFC is engaged in analogical processing (Hobeika et al., 2016; Knowlton, Morrison, Hummel, & Holyoak, 2012; Vartanian, 2012; Wertheim & Ragni, 2018). Consistent with this account, the multivariate patterns within the frontal pole (BA10) and left rlPFC (BA47) were able to discriminate between the valid and invalid trials, further highlighting their involvement in analogy.

We also observed significant clusters preferentially activated during the C:D phase over the A:B phase outside of the rostrolateral PFC, specifically in occipital cortex, the left inferior frontal gyrus, and bilateral posterior parietal cortex. We found that the occipital region did not represent meaningful information to discriminating between valid and invalid analogies (Figure 4.6), consistent with the hypothesis that these activations reflect general difficulty demands, possibly due to participants looking longer at the screen for more difficult or semantically distant stimuli (Green, Fugelsang, Kraemer, Shamosh, & Dunbar, 2006; Green et al., 2010, 2012).

Among the remaining regions, the left IFG and left inferior parietal ROIs did appear to also represent relevant information to discriminating between C:D and C':D'. It has been recently reported that the left rlPFC and posterior parietal cortex are strongly functionally

associated during reasoning tasks, but not known to have a direct anatomical connection (Vendetti & Bunge, 2014; Wendelken et al., 2016). As the left IFG has been implicated in "representing" and "retrieving" relational information (Krawczyk, 2012), it is possible that the superior longitudinal fasciculus, which connects the IFG to the inferior parietal lobe (Burks et al., 2017) facilitates the transfer of relational information from the inferior parietal cortex to rlPFC through the IFG.

On the other hand, activity in the left IFG and posterior parietal regions have also been shown to reflect task general (e.g. working memory and difficulty-related) demands (Fedorenko, Duncan, & Kanwisher, 2012; Fedorenko et al., 2013; Fedorenko & Varley, 2016). While the lack of a working memory control task limits our ability to discuss the task-generality or specificity of the IFG, we see a gradient of information representation in the posterior parietal cortex (PPC). The classification results indicate that while the superior parietal lobe did not meaningfully encode whether the participant was considering a valid or invalid trial, the inferior regions (aSMG, pSMG) could, supporting results from a recent meta analysis that implicate the inferior PPC in reasoning across domains (Wendelken, 2015).

The neural computations underlying analogy By testing the trial-by-trial predictions of these models, we were able to characterize what information was being computed during the two stages. The RSA analysis on relation representation indicates that the left frontal cortex and posterior temporal cortex represent the semantic attributes corresponding to the two words in the word pair, consistent with prior accounts of these regions (Vendetti & Bunge, 2014; Wendelken et al., 2016). During this stage, we did not observe correlations within the rostrolateral PFC, which is consistent with the account that the rlPFC is primarily

engaged when evaluating second order relations (Wendelken, 2015).

We did observe significant correlations with the relational models (Word2vec-diff and BART) in left posterior parietal cortex (SPL, aSMG, pSMG) and near the temporo-parietal junction (AG, toMTG), consistent with prior studies and meta-analyses that directly implicate the posterior parietal cortex in relational reasoning (see Wendelken, 2015 for a review).

According to our RSA analysis on the relation representation stage, the left frontal cortex encoded both semantic (the Word2vec-concat) and higher order (BART) relational information between the two constituent word pairs. However, the voxelwise analysis provided further insight as to the computations necessary for the brain to process analogies. By correlating trial-by-trial relational similarity derived from our computational models with voxel activations, we found that BART-derived relational similarity successfully predicted neural activations in the same frontal regions, while the other two models, Word2vec-concat and Word2vvec-diff, could not. A small cluster in parietal cortex was also correlated with BART-derived relational similarity (and not other models), supporting the hypothesis that relations are formed in the parietal cortex and then transmitted to rlPFC for higher-order reasoning operations.

A neurocomputational model of analogical reasoning Taken together, our results support a recent proposal that subdivisions the left posterior parietal cortex (PPC) are critical to relational reasoning in general (Wendelken, 2015), rather than only visuospatial processing (Krawczyk, 2012; Hobeika et al., 2016), while others are specialized for the many other operations (Duncan, 2010).

By analyzing these two subprocesses via our sequential design, our results also support a coarse, but neurobiologically plausible sequence of events for analogical reasoning. We find that during relational representation, the left parietal cortex computes first order relations between constituents, which are then represented in the fronto-temporal "semantic network". During comparison or integration, the rlPFC computes second order relational judgments over these representations. Despite the lack of direct anatomical connection between rlPFC and the left PPC, it has been recently reported that these two regions are functionally connected, and that this connectivity predicts reasoning ability during development (Vendetti & Bunge, 2014; Wendelken et al., 2016, 2017). Additionally, this sequence of events is corroborated by lesion studies of general reasoning. In one case, when comparing stroke patients with damage to rlPFC or PPC, only patients with damage in PPC were significantly impaired on a transitive inference task (Waechter, Goel, Raymont, Kruger, & Grafman, 2013).

Finally, we also provide an account for the computations carried out in PPC and rlPFC during analogical reasoning. BART captures the higher order process involved in forming an abstract relation. We observed that BART predictions were correlated with ROIs in the posterior parietal cortex, providing a computational account for how relations are formed. Interestingly, in an exploratory (i.e. voxelwise) approach, we found that the rlPFC and other previously implicated frontal regions tracked relational similarity scores derived from BART representations, while other models were not sufficient to explain voxel activity in these regions.

4.4 Methods

Participants 16 participants (8 female) were recruited at the University of California, Los Angeles (UCLA) through a flyer distributed in the Psychology department. Participants signed informed consent prior to the experimental session, and were paid $50 for their participation in the 1-hour study, in compliance with the procedures accepted by the local institutional review board (IRB).

Stimuli The experiment was administered using PsychoPy2 (Peirce, 2009). The basic stimuli were a set of analogy problems constructed on the basis of prototypicality norms for examples of two-word semantic relations (Jurgens et al., 2012). These norms were in turn based on a taxonomy of semantic relations described by (Bejar et al., 1991). The full norms include examples of word pairs instantiating ten high-level categories (e.g., Similar, Contrast, Cause-Purpose), each with five to ten more specific subcategories, for a total of 79 distinct subcategories. For the present study, we focused on three high-level categories (those listed above), with three subcategories of each for a total of nine subcategories.

For each subcategory, the (Jurgens et al., 2012) norms rank roughly 35 word pairs from most to least illustrative, based on human ratings. Each pair is assigned a "prototypicality score" reflecting its level of goodness as an example of the specified relation. Because we wished to create analogy problems with clear correct answers, for each subcategory we selected 16 word pairs from among the most highly rated examples. In making this selection we avoided duplicate pairs that were simple reversals (e.g., happy-sad and sad-happy), in such cases choosing just the pair with the higher rank. Pairs that included conspicuously

long or low-frequency words were also excluded. Because for some subcategories it proved difficult to identify 16 pairs that passed our selection criteria, we also included some pairs that (Jurgens et al., 2012) had used as "seed" examples to elicit word pairs from humans. These were considered excellent examples (most taken from (Bejar et al., 1991); hence we assigned them a prototypicality score equal to that of the most highly ranked example for the subcategory.

Counterbalancing to form analogy problems Using the 144 (16 examples x 9 subcategories) distinct word pairs selected as described above, we formed pairs of pairs to create verbal analogy problems in the form *A:B* :: *C:D* (valid) or else *A:B* :: C':D' (invalid), where all pairs were drawn from the pool of 144. For the invalid pairs, the C':D' pair was drawn from a different higher-order relation category than was A:B. We avoided creating invalid items using different subcategories within the same superordinate category (e.g., subcategories Contrary and Pseudoantonym, both subtypes of Contrast) because pilot work suggested that such "near-miss" problems would lead to excessive errors in a task involving time pressure. At the same time, C':D' pairs always instantiated a natural semantic relation (rather than being semantically anomalous), forcing participants to consider the paired relations carefully in judging validity of the analogies.

Counterbalancing was used to create four complete sets of analogy problems. To form an individual set, for each of the nine subcategories, eight of the 16 pairs were assigned to the *A:B* role and four to the *C:D* role. The remaining four pairs were assigned to the C':D' role associated with *A:B* pairs for four of the six subcategories representing the two remaining higher-order categories. Assignments to the *C:D* role were random subject to the above

restriction. Subject to all of the above restrictions, specific 4-term analogy problems were created by random pairing of word pairs. Each set thus consisted of 72 analogy problems (9 subcategories x 8 problems each). For each subcategory, four problems were valid and four were invalid. Within a set of 72 problems, each of the 144 word pairs occurred twice in the *A:B* role and once in each of the *C:D* and C':D' roles.

The same procedure was used to create a total of four sets, each with 72 problems distributed as described above. Across all four sets, each of the 144 word pairs appeared in each role with the same proportions (i.e., twice as often as *A:B* than as *C:D* or C'D'). The four sets, with a total of 288 problems (4 sets x 72 problems each), were treated as four blocks administered to each participant. The procedure for problem generation ensured that any individual analogy problem occurred only once in the set of 288 problems. The order of problems was randomized within each block, and the order of the four blocks was counterbalanced across participants. The overall aim of this procedure for problem creation was to ensure that data analyses could be based on neural patterns associated with each of the 16 word pairs representing each of the nine relational categories (144 pairs in total), in each of the three possible roles (A:B, C:D, C':D'), while avoiding any confounding between specific pairs and roles. Finally, each of these four sets was further split into two sets of 36 for subject convenience.

Procedure On each trial (see Figure 4.1), participants were first shown the *A:B* word pair for 2s, then the *C:D* pair for 2s (with an average .5s jitter in between). The words "yes" or "no" then appeared on the left and right of the screen, indicating the assignment of two response buttons used to indicate whether or not the two pairs represented the same

relation. Critically, the assignment of "yes" and "no" buttons was randomly varied, ensuring that participants could not begin planning a motor response during the earlier phases of the trial.

Overall, subjects completed the task with an average accuracy of 82% (SD: 6.8%) and response time of 981ms (SD: 30.7).

fMRI Data Acquisition Data were acquired on a 3 Tesla Siemens Prisma Magnetic Resonance Imaging (MRI) scanner at the Staglin IMHRO Center for Cognitive Neuroscience at UCLA. Structural data were acquired using a T1-weighted sequence (MPRAGE, TR = 1,900 ms, TE = 2.26 ms, voxel size 1 mm^3 isovoxel). Blood oxygenation level dependent (BOLD) data were acquired with a T2*-weighted Gradient Recall Echo sequence (TR = 1,000 ms, TE = 37 ms, 60 interleaved slices (2mm gap), voxel size 2x2x2 mm, 6x multiband acceleration).

fMRI Preprocessing Data preprocessing was carried out using FSL (Smith et al., 2004). Prior to univariate analyses, data underwent preprocessing steps including motion correction, slice-timing correction (using Fourier-space time-series phase-shifting), spatial smoothing using a Gaussian kernel of 5 mm full-width half-max, and highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with s=50.0s). Data from each individual run were analyzed employing a univariate general linear model approach (Monti, 2011) inclusive of a pre-whitening correction for autocorrelation.

Spatial smoothing was omitted from the above preprocessing steps for classification and representational similarity analysis in order to preserve spatial heterogeneities. Beta-series

(Rissman et al., 2004) parameter estimates were derived using the Least Squares-Separate (LS-S, Mumford et al., 2012 approach, in which all parameter estimates were calculated one at a time in an iterative manner using a general linear model.

Univariate Analysis All A:B and C:D trials were coded by main relationship for a univariate analysis using the GLM approach (Monti, 2011) to identify regions engaged in forming semantic relations. The probe condition was included as a condition of non-interest, as well as motion parameters. The GLM analysis was carried out using FSL FEAT (Smith et al., 2004; Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2011). Data from individual runs were aggregated employing a mixed effects model (i.e., employing both the within- and between-subject variance), and using automatic outlier detection.

For the main effects (i.e., A:B – rest, C:D – rest) analysis, Z (Gaussianised t) statistic images were thresholded using a cluster correction of Z>2.3 and a (corrected) cluster significance threshold of p=.05. For the contrast analysis (i.e., C:D – A:B), significance was assessed using FSL randomise with TFCE cluster correction. To correct for "reverse subtractions" (Morcom & Fletcher, 2007), the resulting parametric maps were masked by regions in which C:D had a Z-score greater than 2.3 (i.e., the main effect of C:D phase was significant).

Classification Analyses Classifiers were trained to distinguish between the three main relations (similar, contrast, cause-purpose), and were evaluated using a leave-one-run-out cross validation approach (see Etzel & Braver, 2013). For each participant, two such classifications were run: one on the A:B trials and one on the C:D (valid) trials. C:D (valid) trials were selected so that the relation representations during the C:D phase would not

be confounded by additional cognitive operations associated with processing an inconsistent relation. We used a 5mm radius sphere and a linear SVM (Abraham et al., 2014; Pedregosa et al., 2011) . Statistical significance was assessed using FSL randomise with TFCE cluster correction (Smith & Nichols, 2009; Winkler et al., 2014).

Representational Similarity Analysis Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008; Kriegeskorte & Kievit, 2013) was used to characterize the similarities of neural responses across pairs. RSA characterizes the representation in a brain region by a representational dissimilarity matrix (RDM), and compares this empirical matrix with a theoretical model. An RDM is a square symmetric matrix, with each entry referring to the dissimilarity between the activity patterns associated with two trials (e.g., entry (1,2) would represent the dissimilarity between activity patterns on trial 1 and trial 2). Procedurally, each element of the RDM is calculated as 1 minus the Pearson correlation between the beta-series for each pair of trials (Carota et al., 2017; Nili et al., 2014).

Hypothesis models were manually generated to reflect idealized RDMs expected given a theoretical representational space. We generated theoretical RDMs from each of the three neurocomputational models. Each model uses a different calculation to yield a feature vector characterizing a word pair, but the RDM was calculated in the same way for all models as the cosine distance between word-pair representations.

RDMs and hypothesis models were compared by calculating a "second-order similarity" (Nili et al., 2014) which was the Spearman correlation coefficient between the two matrices. All analyses were carried out using Python, making extensive use of the machine learning packages Scikit-learn (Pedregosa et al., 2011) and NiLearn (Abraham et al., 2014). Data

and Analysis code is available online.

ROI selection Within the left prefrontal cortex, we selected the following regions: the left rostrolateral prefrontal cortex (rlPFC- BAs 10 and 47) , the left inferior frontal gyrus (BA 44, 45, 47), and the middle frontal gyrus (BA46). The left temporal cortex was segmented into the following ROIs: the temporal pole (TP), anterior superior and inferior temporal gyrus (aSTG, aITG), and the posterior superior, middle, and temporal gyrus (pSTG, pMTG, pITG). At the temporo-parieto-occipital junction, we selected the middle temporal gyrus (toMTG), and the angular gyrus (AG). Finally, we selected the left superior and inferior parietal lobe (SPL, aSMG, pSMG) in the in the parietal cortex. All ROIs were defined using the Harvard-Oxford Atlas (Desikan et al., 2006) except for the occipital, which was selected from the AAL (Tzourio-Mazoyer et al., 2002) . ROIs were selected mostly based on two reviews (Binder et al., 2009; Krawczyk, 2012). As a control, cerebral spinal fluid (CSF), a region that would not plausibly be involved in processing of abstract semantic relations, was included as a "region of disinterest". The CSF ROI was manually drawn for each subject.

Model details All quantitative models used to create theoretical RDMs for the RSA analysis were based directly (Word2vec-concat, Word2vec-diff) or indirectly (BART) on the outputs of a machine learning model, Word2vec (Mikolov et al., 2013). This model takes a large text corpus (Google News) as input, examines distributional statistics relating each word to neighboring words in sentences (local context), and outputs a modular vector representation for each individual word, termed a word embedding. Word2vec vectors of length 300 were obtained for all words used in the present study. Word2vec-concat (the concatenation of the

vectors for the two words in a pair) and Word2vec-diff (the difference vector derived from the two individual vectors) were calculated and used to create theoretical RDMs.

The BART model (*Bayesian Analogy with Relational Transformations*; Lu et al., 2012) was applied to learn specific relations between word pairs. BART takes as inputs pairs of positive and negative examples of a given relation, where each pair is represented by the concatenation of the Word2vec vector for each word. For example, a vector formed by concatenating the individual vectors for love and hate would constitute a positive example of the antonymy relation, but a negative example of the category membership relation. The model used supervised learning with 20 positive examples and a fixed set of 64-74 negative instances (the top example for each relation from each general category other than that of the target relation) to form weight distributions representing each of the 79 relations in the Jurgens et al., 2012. For each word pair used in the study, these learned weights were used to calculate the posterior probability that the pair instantiated each of the 79 learned relations. The vector of length 79 formed by these posterior probabilities represented the specific relation between the two words in the pair. These vectors were used to create BART's theoretical RDMs.

Similar		
Synonym	Attribute Similarity	Change
big:large	book:magazine	acceleration:speed
boat:ship	chair:sofa	darken:color
car:auto	fence:hedge	death:population
careful:cautious	hill:mountain	dim:light

couch:sofa	house:tent	discount:price
cute:adorable	ladder:stairs	flood:water
house:home	paper:parchment	force:pressure
kid:child	pencil:pen	heat:temperature
make:manufacture	picture:drawing	inflation:price
option:choice	pillow:cushion	lower:volume
pants:trousers	rake:fork	raise:salary
pretty:beautiful	shovel:spoon	rise:tide
raise:elevate	stairs:ladder	shorten:distance
run:sprint	sword:knife	soften:voice
spin:twirl	table:desk	speed:movement
teach:instruct	wagon:trailer	terror:fear

Contrast

Contrary	Directional	Pseudoantonym
accept:reject	ahead:behind	bright:dull
big:small	below:above	day:evening
black:white	climb:descend	enthusiastic:lazy
bright:dark	east:west	fun:boring
dark:light	forward:backward	funny:serious
difficult:easy	front:back	good:wrong
dirty:clean	high:low	high:down
fast:slow	in:out	just:unfair

fat:thin	interior:exterior	loud:discreet
good:bad	left:right	low:up
hot:cold	north:south	majority:small
old:young	rise:sink	obey:protest
pretty:ugly	start:finish	powerful:meek
rich:poor	top:bottom	right:bad
tall:short	under:over	smiling:sad
warm:cool	up:down	witty:dumb

Cause/Purpose

Cause:effect	Cause:compensatory action	Activity:goal
accident:damage	anger:yell	advertise:promote
bath:cleanliness	coldness:shiver	bathe:clean
disease:sickness	danger:flee	breathe:live
exercise:fitness	dirtiness:bathe	burnish:shine
explosion:damage	dirty:bathe	cook:eat
fire:burns	fright:scream	drink:hydrate
germs:sickness	happiness:smile	exercise:healthy
heater:warmth	heat:sweat	flee:escape
illness:discomfort	hunger:eat	ignite:burn
injury:pain	loneliness:socialize	read:learn
joke:laughter	nervousness:sweat	sleep:rest
loss:grief	sadness:cry	speak:express

repetition:boredom	sickness:medicate	study:learn
stimulus:response	thirst:drink	trim:shorten
tragedy:tears	thirsty:drink	wash:clean
workout:sweat	tiredness:rest	work:earn

Table 4.1: Word pairs

CHAPTER 5

Incorporating priors from computational models into neuroimaging encoding methods via Tikhonov regularization

5.1 Introduction

Voxelwise encoding methods (Naselaris et al., 2011) have recently gained popularity as a framework for linking cognitive neuroscience with advances in computational modeling. They have been used to relate image representations in the brain with psychophysics and machine learning inspired models (Kay, Naselaris, Prenger, & Gallant, 2008) and with computational linguistics (Mitchell et al., 2008; Huth, Nishimoto, Vu, & Gallant, 2012; Pereira, Botvinick, & Detre, 2013), and also have been used to relate the neural representation of natural language with computational linguistics (Huth et al., 2016; Wehbe et al., 2014; de Heer et al., 2017). These methods are among the leading approaches behind the recent push for characterizing how stimulus features are represented in neural space, as opposed to simply localizing cognitive operations (Diedrichsen & Kriegeskorte, 2017), allowing for a more comprehensive functional description of brain regions than conventional approaches (Naselaris et al., 2011) within the same simple framework as parametric univariate analysis (Posner &

Deheane, 1994).

What makes these approaches so appealing is the fact that by using feature spaces derived from computational models, as opposed to indicator columns coding for when a given condition is on or at different levels of intensity (Huth et al., 2012, 2016), the encoding framework uses brain data to evaluate the (neural) plausibility of computational models of human behavior and adjudicate between different proposals on the basis of how well they fit neural data (Huth et al., 2016). Furthermore, since the feature spaces can be flexibly specified by the researcher to best capture the most central properties of a model (Cohen et al., 2017), these models allow for naturalistic experimental conditions (Haxby et al., 2011; Wehbe et al., 2014; Nishimoto & Nishida, 2016; Huth et al., 2016; de Heer et al., 2017) as well as more "classical" controlled experimental designs (Mitchell et al., 2008; Kay et al., 2008; Huth et al., 2012). However, encoding methods have been relatively slow to catch on compared to others (for example, (Haxby et al., 2014)) for one main reason: in order to use high dimensional feature spaces, enough observations (i.e., timepoints) are needed to properly fit the model.

One particular tool that addresses this issue is *regularization*. In the GLM framework, regularization is typically used to fit ill-posed or rank-deficient problems, generally when there are as many, if not more predictors than observations. As will be shown below, the most common form of regularization in neuroimaging analysis, ridge regression (Hoerl & Kennard, 1970), operates by limiting the overall magnitude of the estimated parameters. When using this technique, all parameter estimates are "penalized", (i.e. pushed towards zero), so less meaningful features will contribute less to prediction.

A recent development in the encoding framework is the renewed emphasis on the idea that regularized solutions to the General Linear Model (GLM) can be interpreted as Bayesian inference (Diedrichsen & Kriegeskorte, 2017). As will be explained below, the regularization technique is mathematically identical to imposing a prior belief upon the relationship between our predictors. In the case of ridge regression, the implicit prior assumption is that all features ought to be equally penalized (i.e., they are all independent from one another). While this is a general-purpose approach that has been used to great effect, there is no reason why we cannot apply more penalization to some predictors and less to others. Tikhonov regularization (Tikhonov, 1943, 1963) is the generalization of Ridge regression beyond the "equal penalty" assumption, providing a mechanism for incorporating arbitrary penalty configurations, but at the cost of losing an analytic solution (which allows for efficient computing). From a computational modeling standpoint, this implies that not only does the encoding approach allow us to incorporate rich feature spaces that represent the stimuli, but via Tikhonov regularization we are also able to include information on how those features ought to be related to each other (Figure 5.1.

In this paper, we review Tikhonov regularization (Tikhonov, 1943, 1963) within the Bayesian framework (as shown in Vogel, 2002). Then, we describe an efficient implementation for Tikhonov regularization, which rotates the data such that existing GLM solvers can be used. We validate these two points first using a simulation study and then on a case study connecting a deep-learning derived feature set (Mikolov et al., 2013) an existing fMRI dataset (Wehbe et al., 2014).

a. OLS solution	b. Ridge solution	c. Tikhonov solution

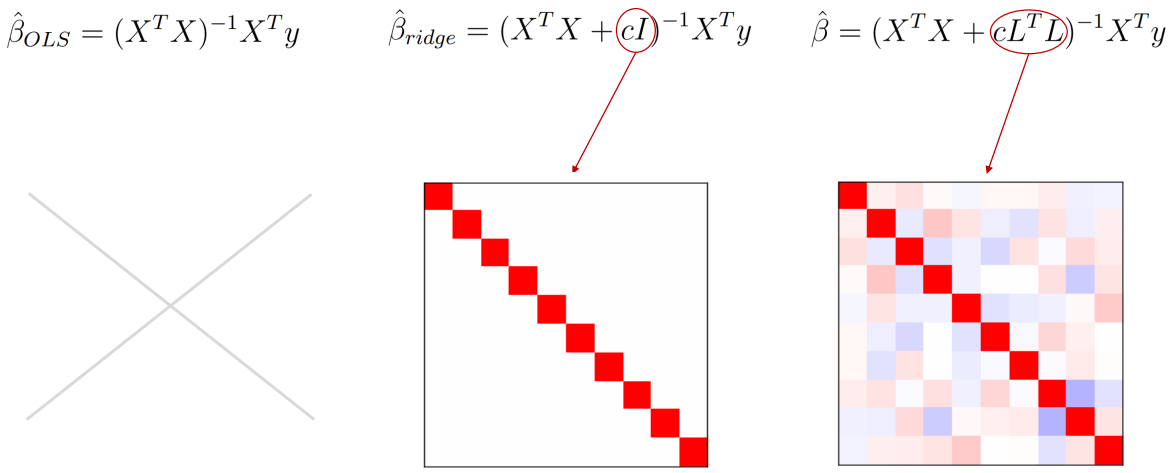

Figure 5.1: Left: OLS does not assume any prior distribution on parameter estimates. Middle: Ridge regression assumes that parameter estimates follow multivariate normal (with identity covariance matrix). Right: Tikhonov regression allows for parameter estimates to follow a multivariate normal with arbitrary variance-covariance matrix.

5.1.1 Linear encoding: a probabilistic perspective

The General Linear Model (GLM) based encoding framework assumes a linear relationship between the cognitive feature space (X) and brain signal (Y) (Monti, 2011; Huth et al., 2012, 2016; Kay et al., 2008; Diedrichsen & Kriegeskorte, 2017; Cohen et al., 2017) such that:

$$Y = X\beta + \epsilon \tag{5.1}$$

In this framework, several techniques exist to minimize the estimation error ϵ by finding the best estimate for β, ($\hat{\beta}$) (e.g., Waldorp, 2009), with the most commonly used being the ordinary least squares (OLS) solution:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y \tag{5.2}$$

An additional constraint on the OLS model is that the errors are independent and identically distributed, following a Gaussian distribution such that:

$$\epsilon \sim Normal(0, \sigma^2) \tag{5.3}$$

Given the constraint of Gaussian error, the linear regression problem can also be formulated probabilistically, as follows:

$$p(y|X, \beta, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(1/2\sigma^2 (y - X\beta)^T (y - X\beta)\right) \tag{5.4}$$

With solution (by optimizing beta over the log-likelihood, see details in the Appendix):

$$\hat{\beta}_{MLE} = (X^T X)^{-1} X^T y \tag{5.5}$$

Thus the Ordinary Least Squares solution for weight estimation (eq. 5.2) can be understood as the maximum likelihood estimator (MLE; eq. 5.5), the best estimator given only our data assuming errors are independent and identically distributed (iid).

This probabilistic re-framing of the linear model approach is what allows the integration of *a priori* knowledge from experimental data (Mitchell et al., 2008) or computational models (Kay et al., 2008; Huth et al., 2016) into the estimation. Thus, the relationship between the contributions of these features $P(\beta)$ can be known or measured. The use of Bayes rule allows us to convert the MLE solution to Maximum a Posteriori (MAP) inference.

We can use Bayes rule to reveal the form of the posterior:

$$p(\beta|y, X, \sigma^2) \propto (y|X, \beta, \sigma^2)p(\beta) = p(y|X, \beta, \sigma^2)p(\beta) \tag{5.6}$$

While the MAP estimator does not have a general analytic solution, conjugate priors can be employed to derive analytic estimators. Specifically, the likelihood function (equation 5.5) follows a Gaussian distribution, which allows us to choose the conjugate prior for β to be a Gaussian:

$$p(\beta) \propto (\tau^2)^{-\frac{p}{2}} \exp\left(-\frac{1}{2\tau^2}(\beta - \mu)^T(\Lambda)(\beta - \mu)\right) \tag{5.7}$$

$$\beta \sim Normal(\mu, \tau^2\Lambda^{-1}) \tag{5.8}$$

Resulting in the full multiplication:

$$p(\beta|y, X, \sigma^2) \propto p(y|X, \beta, \sigma^2)p(\beta) \tag{5.9}$$

By optimizing the log of the above function w.r.t. β, the maximum a posteriori estimate is: (see appendix for proof)

$$\hat{\beta}_{MAP} = (X^T X + c\Lambda)^{-1}(X^T y + c\Lambda\mu) \tag{5.10}$$

The identical estimator can be derived by optimizing the least squared error function, subject to a magnitude constraint:

$$||L\beta||^2 < c \tag{5.11}$$

resulting in

$$\hat{\beta} = (X^T X + cL^T L)^{-1} X^T y \tag{5.12}$$

By setting $L^T L$ to be the inverse variance-covariance matrix (Λ) from equation (5.10) , we see that Tikhonov regularization (Tikhonov, 1963, 1943) is equivalent to the MAP extension to the general linear model.

The Gaussian prior $P(\beta)$ can be characterized by two parameters. In most cases the data are de-meaned or standardized prior to analysis, so prior mean μ is 0. This exposes a mechanism for incorporating the prior—we can choose the regularization matrix such that it

is the inverse of the desired covariance matrix. We are left with the free parameter c, which reflects the how much weight to place on prior knowledge.

We also can the special case when the covariance matrix Λ is the Identity matrix, such that

$$\beta \sim Normal(0, I) \tag{5.13}$$

The solution is identical to that of Ridge (L^2-regularized) Regression which, from a Bayesian perspective, assumes the prior that all features are independent of one another.

$$\hat{\beta}_{ridge} = (X^T X + cI)^{-1} X^T y \tag{5.14}$$

As outlined in Diedrichsen & Kriegeskorte, 2017, ridge regression can be interpreted as Bayesian inference in which the prior distribution is multivariate Gaussian with a diagonal (identity) variance-covariance matrix. Therefore, ridge regression carries the assumption that all features are independent of one another and thus are (equally) penalized. Model fit is optimized by tuning the scaling parameter c, which controls the compromise between the prior information and likelihood function.

Given enough training data, ridge regression has been used to successfully map rich feature spaces to neuroimaging data (Pereira et al., 2013), and to run encoding models using a finite impulse response hemodynamic model (which increases the number of predictors by a factor of the number of finite impulses assumed), (Wehbe et al., 2014; Huth et al., 2016). However, from a Bayesian modeling perspective this approach is suboptimal, because the prior distribution is mis-specified.

By incorporating an independently derived covariance structure (Λ^{-1}, above), we can construct a model that contains *a priori* knowledge of the underlying feature space (rather than assuming independence of features) for a better fit. For example, when fitting an FIR model instead of a hemodynamic response, temporal correlations can be incorporated via the prior distribution. At a higher level, in computational studies on human cognition (usually in psychophysics), we can explicitly define the covariance structure of features.

I propose in the following section an efficient method for computing the Tikhonov solution given a covariance structure of interest.

5.2 Tikhonov regularization in practice

To efficiently apply Tikhonov regularization in the "massive univariate approach" (Monti, 2011) on neuroimaging data, it is necessary to have an efficient algorithm. As outlined in (Stout, Kalivas, & Héberger, 2007), it is possible to rotate the feature space with linear transformation L such that all features are independent (i.e., following equation 5.13 and satisfying the assumptions for ridge regression). With this rotation, it is then trivial to fit the model using an existing ridge regression implementation, then rotate the results back to the original space. Below, I propose a procedure for finding this rotation matrix L.

Formally, given some symmetric and invertible covariance structure Σ we want to find the Tikhonov matrix L such that

$$L^T L = \Sigma^{-1} \tag{5.15}$$

We can take advantage of the singular value decomposition (SVD) of the inverse variance-covariance matrix such that:

$$\Sigma^{-1} = USV^T \tag{5.16}$$

Where S is a matrix with elements only on the diagonal, and U and V^T are the left and right eigenvectors, respectively. In the case where Σ^{-1} is symmetric (as is the case here), we can select $L = DV^T$ where D is the square root of every element in S. With the selection of L, we are able to rotate the Tikhonov regularization problem such that the regularization matrix is the identity matrix (transforming into the "standard form" L^2-regularization (Stout et al., 2007)) in order for efficient computation. That is, when Tikhonov matrix L is known, any problem

$$argmin_\beta(X\beta - y) + ||L\beta||^2 \tag{5.17}$$

can be converted to ridge regression

$$argmin_{\widetilde{\beta}}(\widetilde{X}\widetilde{\beta} - y) + ||\widetilde{\beta}||^2 \tag{5.18}$$

With the following transformations.

$$\widetilde{X} = XL^{-1}, \widetilde{y} = y \tag{5.19}$$

After which we can recover the desired coefficients with the following:

$$L^{-1}\widetilde{\beta} = \beta \tag{5.20}$$

Where L^{-1} can be the Moore-Penrose pseudoinverse of L.

This transformation makes it possible to harness existing efficient implementations of Ridge regression (e.g., Pedregosa et al., 2011) to quickly apply Tikhonov regularization. We apply this implementation to simulations and an already established dataset (Wehbe et al., 2014) to compare this rotation against Ridge regression in the context of comparing feature spaces in neuroimaging studies.

5.3 Simulation

Regularized linear models are typically been employed when the number of predictors (henceforth "features") approaches or exceeds the number of datapoints (henceforth "samples"). By adding a slight bias via the penalty matrix, regularization provides numerical stability when computing the regression solution. With the understanding that Tikhonov regularization as a refinement of the feature independence assumption in the penalty matrix, we expect that it should outperform other methods in a data-starved setting. The following simulation study was run to verify this behavior.

5.3.1 Method

In order to compare the regression methods, a feature set (X), weights (β), and trial-by-trial noise (ϵ) were randomly generated with the following distributions:

$$X \sim Unif(-1, 1) \tag{5.21}$$

$$\beta \sim Normal(0, \Sigma) \tag{5.22}$$

$$\epsilon \sim Normal(0, I) \tag{5.23}$$

The signal y was created according to the linear model (equation 5.1).

The goal of this simulation was to characterize the performance gains (or losses) arising from the Tikhonov regularization technique as a function of amount of available data. As previously stated, we expected performance to converge to ceiling given enough data points. However, performance should increase faster with a properly specified prior.

GLM estimation was carried out using ordinary least squares (OLS), ridge regression, Tikhonov regression with the properly specified prior (Σ), and Tikhonov regression with an improperly specified prior (Σ^{-1}). The inverse prior was selected as a worst-case scenario for prior specification, to observe how harmful a poor specification could be. The ratio of features to samples was manipulated. In each case, model fitting was performed on 80% of these samples, and the regularization scaling parameter c was fixed at 1. We specifically avoided tuning this parameter because it was possible that in some cases it may be optimal to set c as low as possible (and thus ignore the penalty matrix). Predictive correlation (Huth et al., 2016) between the predicted responses and the actual values of the remaining 20% left-out samples was recorded. For each data scenario, the process was repeated 1000 times.

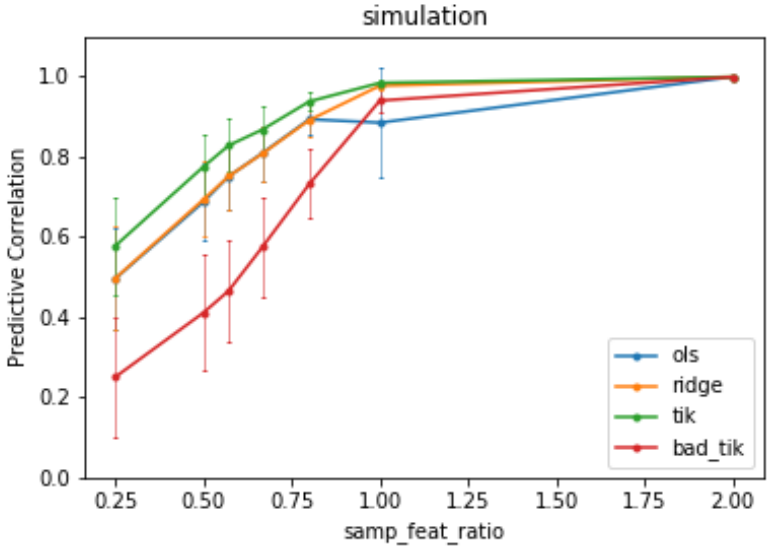

Figure 5.2: Performance when manipulating the number of samples relative to features. As expected, performance increases as more data is available. However, with the properly specified prior Tikhonov regularization outperforms all other methods when there is less data.

5.3.2 Results

Figure 5.2 depicts the result of these simulations. As the ratio of samples to features approaches 1, the regression methods approach maximal predictive correlation. Notably, with fewer samples than features, the proper Tikhonov matrix outperforms all other methods, while the misspecified Tikhonov regularization matrix harms predictive correlation, in addition to being the slwest to converge..

5.4 Tikhonov encoding in neuroimaging

A key limitation of the simulation study is that we imposed a noise structure consistent with the assumptions of the model. Specifically, our errors were independent and identically distributed, which is seldom the case in neuroimaging analysis (Weisskoff et al., 1993; Woolrich et al., 2001; Bright & Murphy, 2015). Additionally, we were assured that the prior distribution was properly specified. In this section, we leverage an existing dataset (Wehbe et al., 2014) and feature space (Word2vec- Mikolov et al., 2013) to test this approach on real functional neuroimaging data.

Wehbe et al., 2014 acquired fMRI data from 9 subjects at Carnegie Mellon University as they passively read a chapter from Harry Potter and the Sorcerer's Stone (Rowling, 1998). They collected a number of features derived from computational linguistics and applied them (using ridge regression) in an encoding framework (Kay et al., 2008; Huth et al., 2012) to show that features specific to different aspects of language processing (e.g. syntax, semantics) were predictive in different (language related) regions of the brain.

Wehbe and colleagues used a non-negative sparse embedding of word co-occurence frequencies (Murphy, Talukdar, & Mitchell, 2012) as a feature space representing semantic meaning. In this reanalysis of their data, we adopted the feature space defined by Word2Vec, which has been shown to perform similarly to co-occurence derived features (Mikolov et al., 2013) to test the Tikhonov rotation described above. Critically, our selection of Word2Vec vectors allows us to calculate a prior matrix, which we defined as the correlation between the 300 features according to all of the other words in the vocabulary. We restrict our analysis to the regions found to be "semantically relevant" by Wehbe et al., 2014.

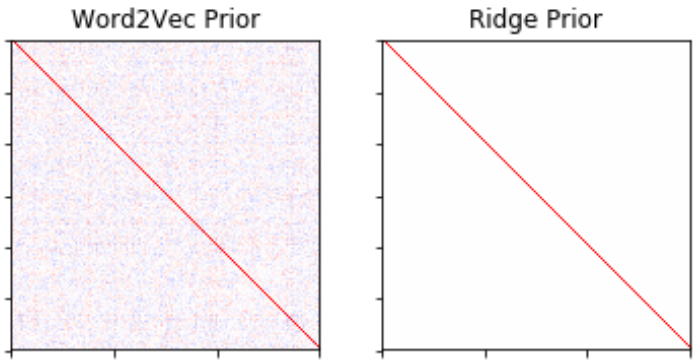

Figure 5.3: Priors derived from Word2vec

5.4.1 Method

As previously described, participants passively read a chapter of Harry Potter and the Sorcerer's Stone. Wehbe and colleagues used a number of features derived from computational linguistics in the encoding framework to identify brain regions involved in processing natural language and found that, for different regions of interest, different feature sets corresponding to linguistic aspects (e.g., syntax, semantics) held the most predictive power.

Design matrices We adapted the classification procedure described by Mitchell et al., 2008 and Wehbe et al., 2014, manipulating the underlying GLM solver and amount of training data. The last 340 timepoints (roughly 20%) were left out as a test set to ensure that the training and testing data could not temporally overlap. Then a voxelwise GLM model was trained on a subset of the remaining timepoints and used to predict the timeseries for the left-out set. To construct the regressors corresponding to word2vec features, the

300-dimensional representations for each word (Mikolov et al., 2013) were convolved with a canonical double-gamma hemodynamic response (HRF) to form the design matrix (NIPY). The predicted timeseries and the left-out timeseries were segmented into non-overlapping sequences of 20 TRs as described in Wehbe et al., 2014, and a classification task was created by comparing the predicted sequence with its corresponding left-out data partition compared with a different randomly selected sequence.

As Word2Vec has been shown to correlate with human semantic judgments (Mikolov et al., 2013), we selected semantically relevant ROIs as determined by Wehbe et al., 2014. Specifically, we restricted our analysis to voxels falling within the inferior frontal gyrus.

We ran the above procedure for varying values of training samples (50-700), across 3 estimation methods: Ordinary Least Squares, Ridge regresion, and Tikhonov regression.

5.4.2 Results

Wehbe et al., 2014 reported up to 74% classification accuracy in this task when using all available linguistic features. With word2vec features alone (which presumably account for only semantics, and not syntax or other thematic relations), the model reaches 65% accuracy.

As the number of training samples decreases, performance diminishes across all methods. However, the regularized regression methods gradually decline in performance, while there is a sharp drop using ordinary least squares. When replicating the clasification task defined by (Wehbe et al., 2014; Mitchell et al., 2008) in which Euclidean distance was used to compare predicted and left-out activity, the magnitude of these distances increased by orders of magnitude as the design matrix was increasingly rank-deficient (not shown).

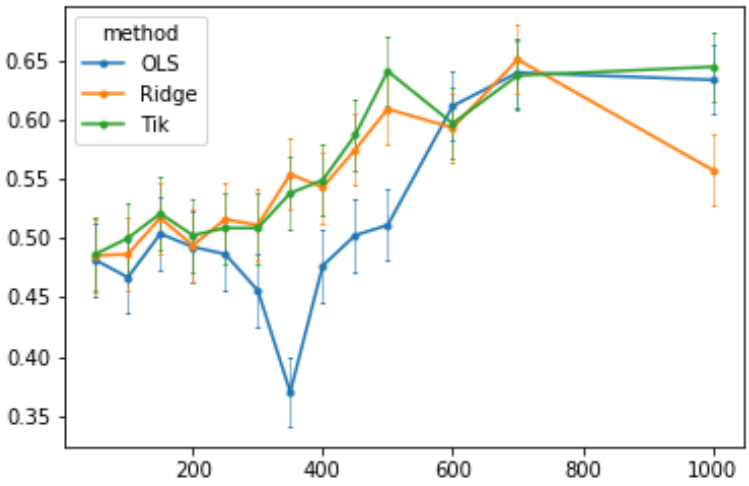

Figure 5.4: Performance when manipulating the number of samples relative to features. As expected, performance increases as more data is available. Predictive correlation is greater for the regularized regression methods while the number of samples is less than the number of features.

5.5 Discussion

This report connects Tikhonov regularization with the Bayesian analysis framework, providing a GLM based mechanism for incorporating prior knowledge into neuroimaging analysis. We accompany this proposal with an efficient implementation in Python to interface with the popular machine learning package Scikit-learn (Pedregosa et al., 2011), and validate the approach on a simulation and neuroimaging analysis.

Simulation As implied by the proof in Appendix C.2 (Vogel, 2002), Tikhonov regularization, when the regularization matrix is properly specified, requires less data to fit a model that reaches maximal predictive power (Figure 5.2). Importantly, model fit appears to increase at a faster rate than simply using ridge regression, which improves at the same rate as the ordinary least squares solution as a function of amount of available training data. This can be attributed to the off-diagonal elements in the regularization matrix, which allow for subsets of variables to be penalized together.

The negative performance associated with a misspecified regularization matrix is a direct consequence of the same off-diagonal elements in the regularization matrix. With two variables that are not truly correlated, artificially linking them together has the effect of pushing the estimator away from the true solution. This underscores the need to carefully measure and incorporate these prior assumptions. If the regression problem is biased in the opposite direction as the true model, the model will at best approximate the OLS solution by reducing the scaling factor c to near zero. However, properly tuning this hyperparameter would cost data, destroying the benefit of needing less data to fit the model.

fMRI case study We applied our implementation to a novel combination of a recent feature space and neuroimaging dataset. Figure 5.4 generally depicts the same pattern of results from the simulation study, in that the regularization methods "learn" to predict the left out test set faster as a function of available training data. While we did not reach the accuracy level (74%) reported by Wehbe et al., 2014, we note that our methods differed in two important ways.

First, Wehbe and colleagues modeled the hemodynamic response function nonparametrically, using a finite impulse response (FIR) type approach (see Pedregosa, Eickenberg, Ciuciu, Thirion, & Gramfort, 2015), while we applied a canonical double gamma hemodynamic response model (Monti, 2011) in order to fit the model. FIR approaches multiply the number of parameters by a factor of t, where t is the number of timepoints used to model the HRF for each voxel, and have been shown to display better predictive accuracy (Marino, Redondo, Luna, Sanchez, & Torres, 2014), but have not been widely adopted due to the propensity to overfit. This is consistent with accounts that the HRF is not consistent across the brain, and simply estimating the amplitude of the double-gamma response function results in temporal information loss. Nonetheless, we chose not to use an FIR approach because the number of required parameters would exceed the number of timepoints, and, as shown in our simulations, we would not have a valid comparison with the OLS method.

Second, Wehbe and colleagues included a number of linguistically relevant features to achieve their optimal prediction. While the Word2vec representation has been shown to capture semantic judgments and associations well (Mikolov et al., 2013), it does not capture any syntactic, relational, or perceptual information (as was included in the full model).

In the applied setting, Tikhonov regression did not outperform standard Ridge regression as it did in the simulation. As shown in figure 5.3, it is possible that the prior matrix as computed by other Word2vec words (which is known to be a very dense representation) was not very different from the prior implied by ridge regression, that is, the 300 dimensions are already nearly independent of one another.

Relation to dimensionality reduction One of the primary goals regularization is to provide numerical stability to the parameter estimate. As previously discussed, the instability in neuroimaging tends to arise from rank-deficiency corresponding to not having enough data. Data-driven dimension reduction techniques (e.g. principal components analysis, independent components analysis) have been used to fit these models at the cost of rotating data into an arbitrary space. Thus, in voxelwise modeling and encoding analyses which aim to interpret the feature spaces being used, this approach might be preferable to these other techniques.

Alternative uses While this study proposes a method for incorporating prior knowledge into the GLM neuroimaging framework, it should be noted that this approach has already been adopted in other imaging domains. Tikhonov regularization has been used for noise reduction in spectrometry studies (Stout & Kalivas, 2007; Stout et al., 2007), in which a regularization matrix functioning as a spectral filter was constructed from known noise characteristics. Also, they have been incorporated into medical image analysis. In a recent study, Lehnert et al., 2018 used Tikhonov regularization in order to increase signal to noise in cardiac MRI by incorporating prior knowledge on signal intensities.

Along those lines, the introduction of Bayesian principles into neuroimaging analysis is not new. Standard structural analysis, particularly segmentations, also incorporate prior expectations on MRI intensity as guided by anatomy (Patenaude, Smith, Kennedy, & Jenkinson, 2011; Jenkinson et al., 2011). However, these algorithms are generally time-inefficient because the posterior distributions typically do not have an analytic form and thus must be approximated using sampling methods. The second contribution of this paper, the rotation procedure, can be applied in these settings to improve development and deployment of new methods, as a rotation of these problems allow for an analytic (ridge) solution rather than relying on an optimization or sampling procedure.

Limitations and Future Directions One limitation that has restricted widespread use of explicit modeling is the need to propose and specify feature spaces for the data. Because model fit serves as the dependent measure, many choices of feature space and prior covariance can result in the same performance (Diedrichsen & Kriegeskorte, 2017). One way to address this is to check when two models actually span the same representational space. This can be determined by comparing the second-moment (covariance) matrix of these feature spaces and priors. Specifically, if two candidate feature spaces' covariances share the same eigenspace, they explain the same intrinsic data and are completely confounded with each other given the available data (see Diedrichsen & Kriegeskorte, 2017 for a detailed explanation and proof).

Beyond the span of the representational spaces themselves, it is also likely that different areas of the brain respond to different feature sets, thereby representing complementary information. Variance partitioning is a recent technique designed to select the best contributors to model fit in different regions of the brain (de Heer et al., 2017).

Finally, as with any study within the Bayesian framework, the choice of prior knowledge is critical. A promising approach may be to measure them in a complementary naturalistic task, as analogous to the hyperalignment technique described in (Haxby et al., 2011, 2014).

Concluding remarks Tikhonov regularization, as presented in this report, can be understood as maximum posterior estimation (MAP) within the GLM framework. It provides a principled method for introducing prior knowledge about a model's predictors into a regression analysis, in addition to providing numerical stability. We provide an efficient framework for implementing this technique by framing it as a rotation into a space where the ridge regression assumption is satisfied. In the context of encoding analysis in neuroimaging research, Tikhonov regularization is a promising framework for bridging the gap between computational modeling and neuroimaging.

CONCLUDING REMARKS

Future directions Among the many theoretical future directions discussed in the individual chapters, several methodological questions persist after the studies in this thesis. These generally reflect preprocessing and free parameter choices that have not been systematically investigated and are not clearly reported in analyses.

While parameter estimation techniques were discussed in Chapter 1, it was apparent when conducting these studies that a more comprehensive follow-up needs to be conducted. While LS-A, LS-S, and the basis function approach to HRF modeling have compared with respect to rapid event-related designs (Mumford et al., 2012), they have not been systematically compared with the FIR method nor have the methods been compared in studies using different organizations for stimuli. For example, in cross-domain tasks (as outlined in Chapter 2), there were several options for coding the nuisance regressor in the LS-S procedure. Along similar lines, it is unclear how these techniques should translate to naturalistic stimuli, despite the field tending towards that direction (Haxby et al., 2014; Huth et al., 2016, for some recent examples).

Similarly, representational similarity analysis is still a developing technique. As it is still under development, there are several new reports showing that different choices of distance metrics provide qualitatively different results (Walther et al., 2016).

Finally, I claim in Chapter 5 that encoding models connect Bayesian modeling with cognitive neuroscience. However, that carries all of the critiques of both analyses. It is still unclear *how* these priors ought to be specified. A follow up in a psychophysical dataset, in

which the prior expectations are clearly known, would be a useful validation of the technique.

Overall, these future directions amount to defining a principled approach to selecting the free parameters for each of the techniques involved in information mapping. Such an overview for each of these methods would be essential for widespread adoption of the methods as well as the reliability of results. The field would benefit from a systematic study of these methods, as well as where to apply them and why.

General Summary The aim of this thesis was to provide a principled approach to bridging the gap between computational modeling in psychology and neuroimaging analysis. The techniques I outline allow researchers to ask probe the underlying computations in previously identified neural correlates of higher cognition.

In Chapter 2, I used canonical multivariate pattern classification (MVPC) in a *cross classification* approach in order to study shared neural representations across language and music. Specifically, we used classification *across* domains to show that structured sequence representation must be shared across language and music within Broca's area, providing a direct link between the extensive prior work on investigating both domains separately.

In Chapter 3, I applied *representational similarity analysis* (RSA) to arbitrate between theories of number representation in the parietal cortex. Because competing theories on rational number representation could explain behavioral data equally well, we turned to neural representations and found that fractions are indeed represented separately from other number types in the posterior parietal cortex.

Chapter 4 represented a combination of classification, RSA, and encoding analyses to

propose a neurocomputational account of relational reasoning. While the neural correlates and computational processes underlying relational reasoning have been extensively explored separately, they have not been directly linked. I applied RSA and encoding-like analyses to let fMRI data arbitrate between three different computational accounts of relation representation and comparison, showing that relational reasoning is indeed a process separate from semantic representation, which is likely carried out in a left fronto-parietal network.

As a theoretical extension to Chapter 4, in Chapter 5 I proposed an extension for encoding analyses (Naselaris et al., 2011) that allows for directly integrating computational models with neuroimaging analysis. I provided an efficient implementation that I will continue to maintain, and validated it in both a simulation study and replication of recent findings on a publicly available dataset.

Taken together, this thesis provide a comprehensive (at the time of writing) list of available techniques for understanding "how", and not just "where" higher cognition happens in the brain.

APPENDIX A

Appendix to Chapter 2

Supplementary Materials for Chiang et al.

Supplementary Tables Table A.1. Activations for the structure versus repeat contrast for language materials. Table A.2. Activations for the structure versus repeat contrast for music materials. Table A.3. Activations for the interaction effect of structure versus repeat and materials.

Supplementary Figures Figure A.1. Univariate analysis. Interaction of the structure versus repeat contrast and materials (i.e., language, music). Figure A.2. ROI mean activity.

Coord (MNI)					
x	y	z	Z	Hem	Anatomical Label (BA)
Frontal					
-44	6	46	5.63	L	Middle Frontal Gyrus (8)
-54	12	16	4.16	L	Precentral Gyrus (44)
-2	10	52	4.04	L	Medial Frontal Gyrus (6)
-58	16	18	3.98	L	Inferior Frontal Gyrus (44)
-4	14	48	3.74	L	Medial Frontal Gyrus (6)
-56	16	12	3.69	L	Inferior Frontal Gyrus (44)

36	-34	62	3.64	R	Precentral Gyrus (4)
-54	26	22	3.62	L	Inferior Frontal Gyrus (45)
6	18	44	3.59	R	Cingulate Gyrus (32)
-4	10	70	3.43	L	Superior Frontal Gyrus (6)
-2	4	70	3.12	L	Superior Frontal Gyrus (6)
-2	14	64	3.1	L	Superior Frontal Gyrus (6)
Temporal					
-52	-46	6	4.78	L	Middle Temporal Gyrus (21)
-58	-30	2	4.49	L	Superior Temporal Gyrus (22)
-52	-50	6	4.19	L	Middle Temporal Gyrus (21)
-62	-48	16	4.17	L	Superior Temporal Gyrus (22)
-50	-34	2	4.11	L	Middle/Superior Temporal Gyrus (22)
Parietal					
40	-44	48	4.84	R	Inferior Parietal Lobule (40)
40	-40	46	4.81	R	Inferior Parietal Lobule (40)
-38	-40	42	4.51	L	Inferior Parietal Lobule (40)
-64	-44	28	4.43	L	Inferior Parietal Lobule (40)
-34	-44	52	3.98	L	Inferior Parietal Lobule (40)
-48	-36	48	3.92	L	Inferior Parietal Lobule (40)
32	-60	50	3.64	R	Superior Parietal Lobule (7)
34	-50	54	3.48	R	Superior Parietal Lobule (7)
22	-62	52	3.43	R	Superior Parietal Lobule (7)

Table A.1: Activations for the structure versus repeat contrast for language materials.

Coord (MNI)					
x	y	z	Z	Hem	Anatomical Label (BA)
Frontal					
-2	10	58	6.8	L	Superior Frontal Gyrus (6)
36	20	4	6.4	R	Insula (13)
-32	24	0	6.2	L	Insula (13/47)
-30	22	6	6.1	L	Insula (13/45)
4	16	48	6	R	Superior Frontal Gyrus (6)
42	22	0	5.8	R	Inferior Frontal Gyrus (47)
50	14	-4	5.6	R	Inferior Frontal Gyrus (47)
-24	-6	56	5.6	L	Middle Frontal Gyrus (6)
-54	-12	4	5.4	L	Inferior Frontal Gyrus (45)
-26	2	56	5.3	L	Middle Frontal Gyrus (6)
46	12	4	5.2	R	Insula (13)
-40	18	-4	5.1	L	Inferior Frontal Gyrus (47)
-4	20	34	5	L	Cingulate Gyrus (32)
8	16	56	4.9	R	Superior Frontal Gyrus (8)
54	10	18	4.8	R	Inferior Frontal Gurys (44)
28	4	58	4.6	R	Middle Frontal Gyrus (6)

10	18	32	4.3	R	Cingulate Gyrus (32)
32	6	64	4.2	R	Middle Frontal Gyrus (6)
10	8	68	4.2	R	Superior Frontal Gyrus (6)
26	10	64	4	R	Middle Frontal Gyrus (6)
28	-6	48	3.2	R	Middle Frontal Gyrus (6)
Parietal					
42	-40	54	7.1	R	Inferior Parietal Lobule (40)
-36	-38	40	6.4	L	Inferior Parietal Lobule (40)
-40	-46	54	6	L	Inferior Parietal Lobule (40)
-2	-60	60	5.4	L	Superior Parietal Lobule (7)
-50	-38	48	5.1	L	Inferior Parietal Lobule (40)
-30	-52	62	5	L	Superior Parietal Lobule (7)
-50	-32	48	4.9	L	Inferior Parietal Lobule (40)
48	-30	48	4.8	R	Inferior Parietal Lobule (40)
24	-66	56	4.3	R	Superior Parietal Lobule (7)
18	-62	46	4	R	Precuneus (7)
56	-30	52	3.9	R	Postcentral Gyrus (40)
40	-34	36	3.8	R	Inferior Parietal Lobule (40)
Subcortical					
34	-66	-24	5.3	R	Lobule VI, Posterior Cerebellum
-24	-72	-26	5	L	Lobule VI, Posterior Cerebellum
10	6	6	4.8	R	Caudate Head

26	-70	-18	4.3	R	Lobule VI, Posterior Cerebellum
2	-56	-22	4.3	R	Lobule V (vermis), Anterior Cerebellum

Table A.2: Activations for the structure versus repeat contrast for music materials.

Coord (MNI)					
x	y	z	Z	Hem	Anatomical Label (BA)
Interaction:Language v Music					
Temporal					
-64	-44	4	3.87	L	Superior Temoral Gyrus (22)
-50	-34	0	3.15	L	Middle Temporal Gyrus (21)
-64	-46	-6	3.14	L	Middle Temporal Gyrus (21)
-50	-46	6	3.01	L	Middle Temporal Gyrus (21)
-50	-54	10	2.89	L	Superior Temoral Gyrus (22)
Interaction:Music v Language					
Frontal					
4	10	60	4.44	R	Superior Frontal Gyrus (6)
46	26	-4	4.08	R	Inferior Frontal Gyrus (47)
42	24	0	3.86	R	Inferior Frontal Gyrus (47)
52	22	-10	3.78	R	Inferior Frontal Gyrus (47)
48	-6	-4	3.64	R	Insula (13)
52	14	-6	3.57	R	Inferior Frontal Gyrus (47)

8	12	66	3.55	R	Superior Frontal Gyrus (6)
8	12	70	3.52	R	Superior Frontal Gyrus (6)
-6	18	34	3.34	L	Cingulate Gyrus (32)
0	22	52	3.09		Medial Frontal Gyrus (8)
2	18	38	2.88	R	Cingulate Gyrus (32)
Subcortical					
-22	-74	-28	4.07	L	Lobule VI, Posterior Cerebellum
-34	-78	-22	3.58	L	Lobule VI, Posterior Cerebellum
-34	-68	-22	3.57	L	Lobule VI, Posterior Cerebellum
-22	-80	-24	3.52	L	Lobule VI, Posterior Cerebellum

Table A.3: Activations for the interaction effect of the structure versus repeat and materials.

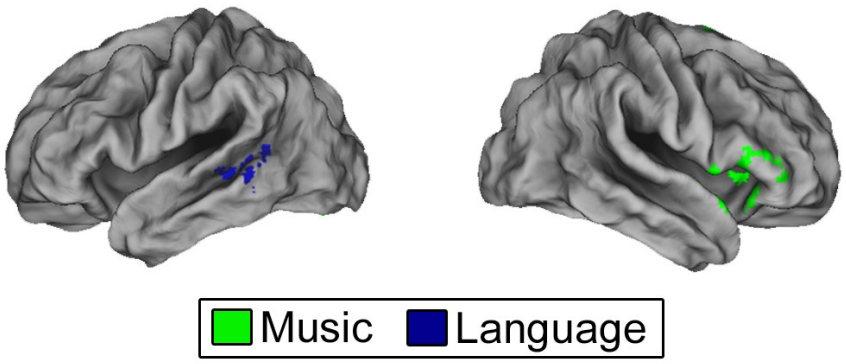

Figure A.1: Interaction of the structure versus repeat contrast and materials (i.e., language, music).

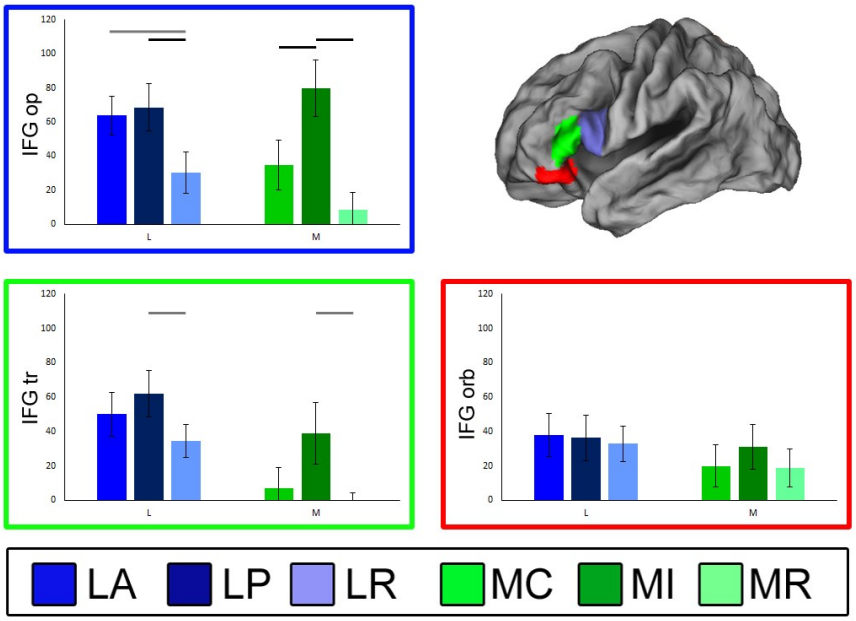

Figure A.2: Mean activity of independently defined subregions of left IFG are presented to demonstrate the pattern of results. The absence of negative values rules out repetition suppression as the primary driver of the significant univariate contrast. (LA, LP, LR: Language Active, Passive, and Repeat; MC, MI, MR: Music Chord, Inversion and Repeat)

APPENDIX B

Appendix to Chapter 4

Coord (MNI)					
x	y	z	Z	Hem	Anatomical Label (BA)
Frontal					
-54	14	2	4.67	L	Inferior Frontal Gyrus (pars opercularis)
-50	10	4	4.66	L	Inferior Frontal Gyrus (pars opercularis)
-4	2	64	5.2	L	Juxtapositional Lobule Cortex (SMA)
0	8	70	4.55		Juxtapositional Lobule Cortex (SMA)
-44	28	-12	4.86	L	Orbitofrontal Cortex
-50	-6	42	4.9	L	Precentral Gyrus
-52	-4	48	4.68	L	Precentral Gyrus
-6	-4	72	3.38	L	Superior Frontal Gyrus
20	-2	14	3.69	R	WM
Occipital					
-34	-88	-4	5.98	L	Lateral Occipital Cortex (inferior division)
-36	-86	-8	5.93	L	Lateral Occipital Cortex (inferior division)
-40	-86	-6	5.81	L	Lateral Occipital Cortex (inferior division)

-46	-70	-18	5.43	L	Lateral Occipital Cortex (inferior division)
48	-64	-22	5.25	R	Occipital Fusiform Gyrus
-34	-92	-2	5.7	L	Occipital Pole
-30	-96	-2	5.48	L	Occipital Pole
22	-92	-12	6.55	R	Occipital Pole
32	-90	-12	6.47	R	Occipital Pole
28	-90	0	6.08	R	Occipital Pole
22	-96	-6	5.5	R	Occipital Pole
24	-94	0	5.33	R	Occipital Pole
Subcortical					
26	2	-14	3.82	R	Amygdala
16	10	4	4.12	R	Caudate
20	10	14	3.83	R	Caudate
-20	10	10	5.63	L	Putamen
24	4	2	4.26	R	Putamen
28	12	-2	3.71	R	Putamen

Table B.1: Local maxima for the A:B trials versus rest contrast.

Coord (MNI)					
x	y	z	Z	Hem	Anatomical Label (BA)
Frontal					

-48	40	4	6.29	L	Frontal Pole
30	40	14	3.33	R	Frontal Pole
-44	16	28	6.43	L	Inferior Frontal Gyrus (pars opercularis)
-44	18	24	5.75	L	Inferior Frontal Gyrus (pars opercularis)
30	30	16	3.83	R	Inferior Frontal Gyrus (pars triangularis)
28	18	12	3.4	R	Insula
38	28	24	4.1	R	Middle Frontal Gyrus
22	18	20	3.71	R	None
28	36	14	3.58	R	None
-42	6	30	6.16	L	Precentral Gyrus
28	-14	52	5.88	R	Precentral Gyrus
30	-12	56	5.62	R	Precentral Gyrus
Occipital					
-36	-84	-8	7.45	L	Lateral Occipital Cortex (inferior division)
-30	-90	-4	7.09	L	Lateral Occipital Cortex (inferior division)
28	-86	-8	7.66	R	Occipital Fusiform Gyrus
22	-92	-12	8.36	R	Occipital Pole
32	-90	-12	7.84	R	Occipital Pole
22	-94	-6	7.75	R	Occipital Pole
Parietal					
-24	-62	40	5.32	L	Lateral Occipital Cortex (superior division)
-28	-60	40	5.15	L	Lateral Occipital Cortex (superior division)

-26	-64	46	5.03	L	Lateral Occipital Cortex (superior division)
-26	-64	34	4.94	L	Lateral Occipital Cortex (superior division)
-26	-50	38	6.62	L	Superior Parietal Lobule
-40	-40	42	5.86	L	Supramarginal Gyrus (anterior division)

Table B.2: Local maxima for the C:D trials versus rest contrast.

Coord (MNI)					
x	y	z	Z	Hem	Anatomical Label (BA)
Frontal					
32	28	14	3.36	R	Frontal Operculum Cortex
-44	40	2	5.21	L	Frontal Pole
32	42	14	3.28	R	Frontal Pole
-42	16	26	5.53	L	Inferior Frontal Gyrus (pars opercularis)
-46	20	22	4.9	L	Inferior Frontal Gyrus (pars opercularis)
36	26	18	3.42	R	Inferior Frontal Gyrus (pars triangularis)
-46	24	24	5.16	L	Middle Frontal Gyrus
40	30	24	3.58	R	Middle Frontal Gyrus
48	34	26	3.45	R	Middle Frontal Gyrus
42	34	20	3.4	R	Middle Frontal Gyrus
-42	8	28	4.98	L	Precentral Gyrus
-44	6	32	4.86	L	Precentral Gyrus

42	-2	14	4.3	R	Central Opercular Cortex
42	2	14	4.08	R	Central Opercular Cortex
48	-4	12	3.51	R	Central Opercular Cortex
40	-2	20	3.27	R	Central Opercular Cortex
38	-20	22	3.52	R	Parietal Operculum Cortex
38	-24	20	3.32	R	Parietal Operculum Cortex
Occipital					
38	-84	-2	5.75	R	Lateral Occipital Cortex (inferior division)
-48	-68	-6	5.72	L	Lateral Occipital Cortex (inferior division)
40	-80	2	5.63	R	Lateral Occipital Cortex (inferior division)
48	-76	4	5.63	R	Lateral Occipital Cortex (inferior division)
20	-96	4	8.45	R	Occipital Pole
16	-92	-14	7.41	R	Occipital Pole
36	-62	54	6.06	R	Lateral Occipital Cortex (superior division)
24	-58	40	5.56	R	Lateral Occipital Cortex (superior division)
-30	-72	48	5.61	L	Lateral Occipital Cortex (superior division)
-18	-70	44	4.46	L	Lateral Occipital Cortex (superior division)
-22	-66	46	4.46	L	Lateral Occipital Cortex (superior division)
Parietal					
2	8	48	5.29	R	Paracingulate Gyrus
40	-28	56	5.59	R	Postcentral Gyrus
32	-28	50	5.34	R	Postcentral Gyrus

30	-12	52	5.17	R	Precentral Gyrus
-26	-50	40	4.94	L	Superior Parietal Lobule
-32	-50	44	4.74	L	Superior Parietal Lobule
-40	-40	42	5.7	L	Supramarginal Gyrus (anterior division)
Temporal					
-62	-42	4	5.67	L	Middle Temporal Gyrus (posterior division)
-52	-48	4	4.48	L	Middle Temporal Gyrus (temporooccipital part)
-64	-36	6	5.4	L	Superior Temporal Gyrus (posterior division)
-48	-32	-2	4.61	L	Superior Temporal Gyrus (posterior division)
-50	-40	8	4.56	L	Superior Temporal Gyrus (posterior division)
-52	-46	10	4.68	L	Supramarginal Gyrus (posterior division)

Table B.3: Local maxima for the C:D - A:B contrast.

APPENDIX C

Appendix to Chapter 5

C.1 MLE OLS Derivation

In this section we repeat the MLE solution to the General Linear Model.

An additional constraint on the OLS model is that the errors are independent and identically distributed, following a Gaussian distribution such that:

$$\epsilon \sim Normal(0, \sigma^2) \tag{C.1}$$

Given the constraint of Gaussian error, the linear regression problem can be formulated probabilistically.

$$p(y|X, \beta, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(1/2\sigma^2 (y - X\beta)^T (y - X\beta)\right) \tag{C.2}$$

$$\frac{d}{d\beta} \log(p(y|X, \beta, \sigma^2)) \tag{C.3}$$

$$= \frac{d}{d\beta} \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) = 0 \tag{C.4}$$

With solution (by optimizing beta over the log-likelihood, see details in the Appendix):

$$\hat{\beta}_{MLE} = (X^T X)^{-1} X^T y \tag{C.5}$$

C.2 Relating Tikhonov regression to MAP estimation

The posterior distribution for β is the following:

$$p(\beta|y, X, \sigma^2) \propto p(y|X, \beta, \sigma^2) p(\beta) \tag{C.6}$$

We optimize this function with respect to β to find the maximum posterior estimate.

$$\frac{d}{d\beta} \log(p(\beta|y, X, \sigma^2)) = \frac{d}{d\beta}(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) - \frac{1}{2\tau^2}(\beta - \mu)^T(\Lambda)(\beta - \mu)) \tag{C.7}$$

$$= \frac{d}{d\beta}(-\frac{1}{2\sigma^2}(y^T y - 2\beta^T X^T y + \beta^T X^T X\beta) - \frac{1}{2\tau^2}(\beta^T \Lambda\beta - 2\beta^T \Lambda\mu + \mu^T \Lambda\mu)) \tag{C.8}$$

$$0 = -\frac{1}{2\sigma^2}(2X^T X\beta - 2X^T y) - \frac{1}{2\tau^2}(2\Lambda\beta - 2\Lambda\mu)X^T X\beta + \frac{\sigma^2}{\tau^2}\Lambda\beta \tag{C.9}$$

$$= X^T y + \frac{\sigma^2}{\tau^2}\Lambda\hat{\beta_{MAP}} \tag{C.10}$$

Solving for β, the maximum a posteriori estimate is:

$$\hat{\beta}_{MAP} = (X^T X + c\Lambda)^{-1}(X^T y + c\Lambda\mu) \tag{C.11}$$

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., … Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Front Neuroinform*, *8*, 14. doi: 10.3389/fninf.2014.00014

Agrawal, P., Stansbury, D., Malik, J., & Gallant, J. (2014). Pixels to voxels: modeling visual representation in the human brain. *arXiv*.

Albers, A. M., Kok, P., Toni, I., Dijkerman, H., & de Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Curr. Biol.*, *23*(15), 1427-31. doi: 10.1016/j.cub.2013.05.065

Alpaydin, E. (2009). *Introduction to machine learning*. MIT Press.

Ansari, D. (2007). Does the parietal cortex distinguish between "10," "ten," and ten dots? *Neuron*, *53*(2), 165–167. doi: 10.1016/j.neuron.2007.01.001

Ayotte, J., Peretz, I., & Brain, H. K. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. *Brain*. doi: 10.1093/brain/awf028

Ayotte, J., Peretz, I., Rousseau, I., Bard, C., & Bojanowski, M. (2000). Patterns of music agnosia associated with middle cerebral artery infarcts. *Brain*, *123*(9), 1926-1938. doi: 10.1093/brain/123.9.1926

Barth, H., La Mont, K., Lipton, J., & Spelke, E. S. (2005). Abstract number and arithmetic in preschool children. *Proc. Natl. Acad. Sci. U.S.A.*, *102*(39), 14116-21. doi: 10.1073/pnas.0505512102

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). *Cognitive and psychometric analysis of analogical problem solving*. Springer-Blackwell New York.

Bek, J., Blades, M., Siegal, M., & Varley, R. (2010). Language and spatial reorientation: evidence from severe aphasia. *J Exp Psychol Learn Mem Cogn*, *36*(3), 646-658. doi: 10.1037/a0018281

Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in Cognitive Sciences*, *17*(2), 89-98. doi: 10.1016/j.tics.2012.12.002

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex (New York, N.Y. : 1991)*, *19*(12), 2767 96. doi: 10.1093/cercor/bhp055

Boeckx, C. (2010). *Language in cognition: Uncovering mental structures and the rules behind them*. Wiley-Blackwell. doi: 10.1002/9781444310047.ch11

Bonato, M., Fabbri, S., Umiltà, C., & Zorzi, M. (2007). The mental representation of numerical fractions: real or integer? *J Exp Psychol Hum Percept Perform*, *33*(6), 1410-9. doi: 10.1037/0096-1523.33.6.1410

Bookheimer, S. (2002). Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. *Annu Rev Neurosci*, *25*, 151-188. doi: 10.1146/annurev.neuro.25.112701.142946

Boynton, G., Engel, S., Glover, G., & Heeger, D. (1996). Linear systems analysis of functional magnetic resonance imaging in human v1. *Journal of Neurosciences*, *16*, 4207 - 4221.

Brannon, E. M. (2002). The development of ordinal numerical knowledge in infancy. *Cognition*, *83*(3), 223-40.

Braver, T., Cohen, J., Nystrom, L., Jonides, J., Smith, E., & Noll, D. (1997). A parametric

study of prefrontal cortex involvement in human working memory. *Neuroimage*, *5*(1), 49-62. doi: 10.1006/nimg.1996.0247

Bright, M. G., & Murphy, K. (2015). Is fMRI "noise" really noise? resting state nuisance regressors remove variance with network structure. *Neuroimage*, *114*, 158-169. doi: 10.1016/j.neuroimage.2015.03.070

Brown, S., Martinez, M. J., & Parsons, L. M. (2006). Music and language side by side in the brain: a PET study of the generation of melodies and sentences. *Eur J Neurosci*, *23*(10), 2791-2803. doi: 10.1111/j.1460-9568.2006.04785.x

Bulthé, J., De Smedt, B., & Op de Beeck, H. P. (2015). Visual number beats abstract numerical magnitude: format-dependent representation of arabic digits and dot patterns in human parietal cortex. *J Cogn Neurosci*, *27*(7), 1376-87.

Bunge, S., Helskog, E., & Neuroimage, W. C. (2009). Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *Neuroimage*.

Bunge, S., Wendelken, C., Badre, D., & Wagner, A. (2005). Analogical reasoning and prefrontal cortex: evidence for separable retrieval and integration mechanisms. *Cerebral cortex*. doi: 10.1093/cercor/bhh126

Burks, J. D., Boettcher, L. B., Conner, A. K., Glenn, C. A., Bonney, P. A., Baker, C. M., … Sughrue, M. E. (2017). White matter connections of the inferior parietal lobule: A study of surgical anatomy. *Brain Behav*, *7*(4), e00640. doi: 10.1002/brb3.640

Buxton, R., Wong, E., & Frank, L. (1998). Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. *Magnetic Resonance in Medicine*, *39*, 855 - 864.

Carota, F., Kriegeskorte, N., Nili, H., & Pulvermüller, F. (2017). Representational similarity mapping of distributional semantics in left inferior frontal, middle temporal, and motor cortex. *Cereb. Cortex*, *27*(1), 294-309. doi: 10.1093/cercor/bhw379

Chang, C., & Lin, C. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*.

Cheung, V. K., Meyer, L., Friederici, A. D., & Koelsch, S. K. (2018). The right inferior frontal gyrus processes nested non-local dependencies in music. *Scientific reports*. doi: 10.1038/s41598-018-22144-9

Choi, H. J., Zilles, K., Mohlberg, H., Schleicher, A., Fink, G. R., Armstrong, E., & Amunts, K. (2006). Cytoarchitectonic identification and probabilistic mapping of two distinct areas within the anterior ventral bank of the human intraparietal sulcus. *J. Comp. Neurol.*, *495*(1), 53-69. doi: 10.1002/cne.20849

Chomsky, N. (1957). *Syntactic structures*. Berlin: Mouton, The Hague.

Chomsky, N. (1983). Noam chomsky s views on the psychology of language and thought. In R. W. Rieber (Ed.), *Dialogues on the psychology of language and thought: conversations with noam chomsky, charles osgood, jean piaget, ulric neisser, and marcel kinsbourne* (p. 33–63). Plenium Press.

Chomsky, N. (1995). *The minimalist program*. The MIT Press.

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., & Gabrieli, J. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *Neuroimage*, *14*(5), 1136-1149. doi: 10.1006/nimg.2001.0922

Christophel, T. B., Cichy, R. M., Hebart, M. N., & Haynes, J. D. (2015). Parietal and early visual cortices encode working memory content across mental transformations.

Neuroimage, *106*, 198-206. doi: 10.1016/j.neuroimage.2014.11.018

Cichy, R. M., Chen, Y., & Haynes, J. D. (2011). Encoding the identity and location of objects in human LOC. *Neuroimage*, *54*(3), 2297-307. doi: 10.1016/j.neuroimage.2010.09.044

Cichy, R. M., Heinzle, J., & Haynes, J. D. (2012). Imagery and perception share cortical representations of content and location. *Cereb. Cortex*, *22*(2), 372-80. doi: 10.1093/cercor/bhr106

Coetzee, J. P., & Monti, M. M. (2018). At the core of reasoning: Dissociating deductive and non-deductive load. *Hum Brain Mapp*, *39*(4), 1850-1861. doi: 10.1002/hbm.23979

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*.

Cohen, J., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., … Willke, T. L. (2017). Computational approaches to fMRI analysis. *Nat Neurosci*, *20*(3), 304-313. doi: 10.1038/nn.4499

Cohen Kadosh, R., Cohen Kadosh, K., Kaas, A., Henik, A., & Goebel, R. (2007). Notation-dependent and -independent representations of numbers in the parietal lobes. *Neuron*, *53*(2), 307-14. doi: 10.1016/j.neuron.2006.12.025

Damarla, S. R., & Just, M. A. (2013). Decoding the representation of numerical values from brain activation patterns. *Hum Brain Mapp*, *34*(10), 2624-34. doi: 10.1002/hbm.22087

de Cavey, J., & Hartsuiker, R. J. (2016). Is there a domain-general cognitive structuring system? evidence from structural priming across music, math, action descriptions, and language. *Cognition*, *146*, 172-184. doi: 10.1016/j.cognition.2015.09.013

Dehaene, S., & Changeux, J. (1993). Development of elementary numerical abilities: a neuronal model. *J Cogn Neurosci*, *5*(4), 390-407. doi: 10.1162/jocn.1993.5.4.390

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cogn Neuropsychol*, *20*(3), 487-506. doi: 10.1080/02643290244000239

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *J. Neurosci.*, *37*(27), 6539-6557. doi: 10.1523/JNEUROSCI.3267-16.2017

Della Rosa, P. A., Catricalà, E., Canini, M., Vigliocco, G., & Cappa, S. F. (2018). The left inferior frontal gyrus: A neural crossroads between abstract and concrete knowledge. *Neuroimage*, *175*, 449-459. doi: 10.1016/j.neuroimage.2018.04.021

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968-80. doi: 10.1016/j.neuroimage.2006.01.021

DeWolf, M., Bassok, M., & Holyoak, K. J. (2015a). Conceptual structure and the procedural affordances of rational numbers: relational reasoning with fractions and decimals. *J Exp Psychol Gen*, *144*(1), 127-50. doi: 10.1037/xge0000034

DeWolf, M., Bassok, M., & Holyoak, K. J. (2015b). From rational numbers to algebra: separable contributions of decimal magnitude and relational understanding of fractions. *J Exp Child Psychol*, *133*, 72-84. doi: 10.1016/j.jecp.2015.01.013

DeWolf, M., Chiang, J. N., Bassok, M., Holyoak, K. J., & Monti, M. M. (2016). Neural representations of magnitude for natural and rational numbers. *Neuroimage*, *141*, 304-312. doi: 10.1016/j.neuroimage.2016.07.052

DeWolf, M., Grounds, M. A., Bassok, M., & Holyoak, K. J. (2014). Magnitude comparison with different types of rational numbers. *J Exp Psychol Hum Percept Perform*, *40*(1),

71-82. doi: 10.1037/a0032916

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *Plos Comput Biol*, *13*(4), e1005508. doi: 10.1371/journal.pcbi.1005508

Diester, I., & Nieder, A. (2007). Semantic associations between signs and numerical categories in the prefrontal cortex. *PLoS Biol.*, *5*(11), e294. doi: 10.1371/journal.pbio.0050294

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences*.

Eger, E., Sterzer, P., Russ, M. O., Giraud, A. L., & Kleinschmidt, A. (2003). A supramodal number representation in human intraparietal cortex. *Neuron*, *37*(4), 719-25.

Etzel, J. A., & Braver, T. S. (2013). MVPA permutation schemes: Permutation testing in the land of Cross-Validation. *IEEE*, 140-143. doi: 10.1109/PRNI.2013.44

Fadiga, L., Craighero, L., & Alessandro, D. (2009). Broca's area in language, action, and music. *Ann N Y Acad Sci*, *1169*, 448-458. doi: 10.1111/j.1749-6632.2009.04582.x

Fazio, L. K., DeWolf, M., & Siegler, R. (2016). Strategy use and strategy choice in fraction magnitude comparison. *J Exp Psychol Learn Mem Cogn*, *42*(1), 1-16. doi: 10.1037/xlm0000153

Fazio, P., Cantagallo, A., Craighero, L., Alessandro, D., Roy, A. C., Pozzo, T., … Fadiga, L. (2009). Encoding of human action in broca's area. *Brain*, *132*(Pt 7), 1980-1988. doi:

10.1093/brain/awp118

Fedorenko, E., Duncan, J., & Kanwisher, N. (2012). Language-selective and domain-general regions lie side by side within broca's area. *Current Biology*, *22*(21), 2059-62. doi: 10.1016/j.cub.2012.09.011

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the …*, *110*(41), 16616-21. doi: 10.1073/pnas.1315235110

Fedorenko, E., Hsieh, P., A, N., S, W., & Kanwisher, N. (2010). New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *Journal of …*.

Fedorenko, E., & Varley, R. (2016). Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Ann Ny Acad Sci*, *1369*(1), 132-153. doi: 10.1111/nyas.13046

Fitch, T. W. (2014). Toward a computational framework for cognitive biology: unifying approaches from cognitive neuroscience and comparative cognition. *Phys Life Rev*, *11*(3), 329-64. doi: 10.1016/j.plrev.2014.04.005

Fitch, T. W., Hauser, M., & Chomsky, N. (2005). The evolution of the language faculty: clarifications and implications. *Cognition*, *97*(2), 179-210; discussion 211-25. doi: 10.1016/j.cognition.2005.02.005

Fitch, T. W., & Martins, M. (2014). Hierarchical processing in music, language, and action: Lashley revisited. *Annals of the New York Academy of …*, *1316*(1), 87-104. doi: 10.1111/nyas.12406

Friederici, A. D. (2011). The brain basis of language processing: from structure to function.

Physiological reviews, *91*(4), 1357-92. doi: 10.1152/physrev.00006.2011

Friederici, A. D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. *Trends in cognitive sciences*, *16*(5), 262 8. doi: 10.1016/j.tics.2012.04.001

Gallistel, C. (1993). A conceptual framework for the study of numerical estimation and arithmetic reasoning in animals. *The development of numerical competence: Animal ….*

Gleitman, L., & Papafragou, A. (2013). The oxford handbook of cognitive psychology. In D. Reisberg (Ed.), (p. 504–523). OUP USA.

Grahn, J., & Brett, M. (2007). Rhythm and beat perception in motor areas of the brain. *Journal of cognitive neuroscience*.

Green, A. E., Fugelsang, J. A., Kraemer, D. J., Shamosh, N. A., & Dunbar, K. N. (2006). Frontopolar cortex mediates abstract integration in analogy. *Brain Res.*, *1096*(1), 125-37. doi: 10.1016/j.brainres.2006.04.024

Green, A. E., Kraemer, D. J., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cereb Cortex*, *20*(1), 70-76. doi: 10.1093/cercor/bhp081

Green, A. E., Kraemer, D. J., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2012). Neural correlates of creativity in analogical reasoning. *J Exp Psychol Learn Mem Cogn*, *38*(2), 264-72. doi: 10.1037/a0025764

Haggard, P. (2008). Human volition: towards a neuroscience of will. *Nature Reviews Neuroscience*. doi: 10.1038/nrn2497

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: a python toolbox for multivariate pattern analysis of fMRI data.

Neuroinformatics, *7*(1), 37-53. doi: 10.1007/s12021-008-9041-y

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Olivetti, E., Fründ, I., Rieger, J. W., … Pollmann, S. (2009). PyMVPA: a unifying approach to the analysis of neuroscientific data. *Front Neuroinform*, *3*, 3. doi: 10.3389/neuro.11.003.2009

Haxby, J., Connolly, A., & Guntupalli, S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annu Rev Neurosci*, *37*(1), 435-456. doi: 10.1146/annurev-neuro-062012-170325

Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425-30.

Haxby, J., Guntupalli, S., Connolly, A., Halchenko, Y., Conroy, B., Gobbini, I., … Ramadge, P. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, *72*(2), 404 16. doi: 10.1016/j.neuron.2011.08.026

Haynes, J. (2015). A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. *Neuron*.

Hebart, M. N., & Baker, C. I. (2017). Deconstructing multivariate decoding for the study of brain function. *Neuroimage*. doi: 10.1016/j.neuroimage.2017.08.005

Hobeika, L., Capucine, D., Garcin, B., Levy, R., & Volle, E. (2016). General and specialized brain correlates for analogical reasoning: A meta-analysis of functional imaging studies. *Human brain mapping*, *37*(5), 1953-1969. doi: 10.1002/hbm.23149

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. doi: 10.1080/00401706.1970.10488634

Holyoak, K. (1978). Comparative judgments with numerical reference points. *Cognitive*

Psychology.

Holyoak, K. (2012). Analogy and relational reasoning. Oxford University Press.

Hugdahl, K., Raichle, M. E., Mitra, A., & Specht, K. (2015). On the existence of a generalized non-specific task-dependent network. *Front Hum Neurosci*, *9*, 430. doi: 10.3389/fnhum.2015.00430

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453-8. doi: 10.1038/nature17637

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210-1224. doi: 10.1016/j.neuron.2012.10.014

Ischebeck, A., Schocke, M., & Delazer, M. (2009). The processing and representation of fractions within the brain: An fMRI investigation. *NeuroImage*.

Jackendoff, R. (2002). *Foundations of language: brain, meaning, grammar, evolution*. Oxford University Press.

Jacob, S., & Nieder, A. (2009a). Notation-independent representation of fractions in the human parietal cortex. *Journal of Neuroscience*. doi: 10.1523/JNEUROSCI.0651-09.2009

Jacob, S., & Nieder, A. (2009b). Tuning to non-symbolic proportions in the human frontoparietal cortex. *European Journal of Neuroscience*. doi: 10.1111/j.1460-9568.2009.06932.x

Jacob, S., Vallentin, D., & Nieder. (2012). Relating magnitudes: the brain's code for proportions. *Trends in cognitive sciences*.

Janata, P. (1995). ERP measures assay the degree of expectancy violation of harmonic contexts in music. *Journal of Cognitive Neuroscience*.

Jenkinson, M., Bannister, P., Brady, J., & Smith, S. (2002). Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*, 825 - 841.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2011). Fsl. *NeuroImage*, *62*(2), 782 90. doi: 10.1016/j.neuroimage.2011.09.015

Jurgens, D., Turney, P., Mohammad, S. M., & Holyoak, K. J. (2012). Semeval-2012 task 2: Measuring degrees of relational similarity. *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*.

Kallai, A., & of Human, J. (2009). A generalized fraction: an entity smaller than one on the mental number line. *Journal of Experimental Psychology: Human ….*

Katz, J., & Pesetsky, D. (2011). The identity thesis for language and music. *Draft published online: lingBuzz/000959*.

Kaufmann, L., Koppelstaetter, F., Delazer, M., & Siedentopf, C. (2005). Neural correlates of distance and congruity effects in a numerical stroop task: an event-related fMRI study. *Neuroimage*.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352. doi: 10.1038/nature06713

Knowlton, B. J., Morrison, R. G., Hummel, J. E., & Holyoak, K. J. (2012). A neurocomputational system for relational reasoning. *Trends in cognitive science*, *16*(7), 373-381. doi: 10.1016/j.tics.2012.06.002

Koelsch, S., Gunter, T., Cramon, D., & Neuroimage, Z. S. (2002). Bach speaks: a cortical

"language-network" serves the processing of music. *Neuroimage*.

Koelsch, S., Gunter, T. C., Wittfoth, M., & Sammler, D. (2005). Interaction between syntax processing in language and in music: an ERP study. *Journal of cognitive neuroscience*, *17*(10), 1565-77. doi: 10.1162/089892905774597290

Koelsch, S., Kasper, E., Sammler, D., & Nature …, S. K. (2004). Music, language and meaning: brain signatures of semantic processing. *Nature …*.

Koelsch, S., Rohrmeier, M., Torrecuso, R., & Jentschke, S. (2013). Processing of hierarchical syntactic structure in music. *Proc. Natl. Acad. Sci. U.S.A.*, *110*(38), 15443-8. doi: 10.1073/pnas.1300272110

Kravitz, D. J., Kriegeskorte, N., & Baker, C. I. (2010). High-level visual object representations are constrained by position. *Cereb. Cortex*, *20*(12), 2916-25. doi: 10.1093/cercor/bhq042

Krawczyk, D. C. (2012). The cognition and neuroscience of relational reasoning. *Brain Res.*, *1428*, 13-23. doi: 10.1016/j.brainres.2010.11.080

Kriegeskorte, N., & Diedrichsen, J. (2016). Inferring brain-computational mechanisms with models of activity measurements. *Phil Trans R Soc B*, *371*(1705), 20160278. doi: 10.1098/rstb.2016.0278

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *PNAS*, *103*(10), 3863 3868. doi: 10.1073/pnas.0600244103

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn Sci*, *17*(8), 401-412. doi: 10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis –

connecting the branches of systems neuroscience. *Frontiers in Systems Neurosciences*, *2*, 1 – 28.

Kriegeskorte, N., Simmons, K. W., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci*, *12*(5), 535–540. doi: 10.1038/nn.2303

Kunert, R., Willems, R. M., Casasanto, D., Patel, A. D., & Hagoort, P. (2015). Music and language syntax interact in broca's area: an fMRI study. *PLoS One*, *10*(11), e0141069.

Lashley, K. (1951). The problem of serial order in behavior. In (p. 112-146). Wiley.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Journal of machine learning research*, 1188-1196.

Lehnert, J., Wübbeler, G., Kolbitsch, C., Chiribiri, A., Coquelin, L., Ebrard, G., … Elster, C. (2018). Pixel-wise quantification of myocardial perfusion using spatial tikhonov regularization. *Phys Med Biol*, *63*(21), 215017. doi: 10.1088/1361-6560/aae758

Lerdahl, F. (2001). *Tonal pitch space*. Oxford University Press.

Lerdahl, F., & Jackendoff, R. (1985). *A generative theory of tonal music*. MIT Press.

Lima, C., Krishnan, S., & Scott, S. (2016). Roles of supplementary motor areas in auditory processing and auditory imagery. *Trends in neurosciences*.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological review*, *119*(3), 617.

Luria, A., Tsvetkova, L., & of the neurological, F. D. (1965). Aphasia in a composer. *Journal of the neurological …*.

Maess, B., Koelsch, S., Gunter, T., & Friederici, A. (2001). Musical syntax is processed in broca's area: an MEG study. *Nat Neurosci*, *4*(5), 540-545. doi: 10.1038/87502

Marin, O., & Perry, D. (1999). Neurological aspects of music perception and performance. *The Psychology of Music (Second Edition).*

Marino, J., Redondo, S., Luna, F., Sanchez, L. M., & Torres, F. (2014). Hemodynamic response in a geographical word naming verbal fluency test. *The Spanish Journal of Psychology*, *17*(e33), 1 7. doi: 10.1017/sjp.2014.31

Maruyama, M., Pallier, C., Jobert, A., Sigman, M., & Dehaene, S. (2012). The cortical representation of simple mathematical expressions. *Neuroimage*, *61*(4), 1444-1460. doi: 10.1016/j.neuroimage.2012.04.020

Meert, G., Grégoire, J., & Noël, M. (2010). Comparing 5/7 and 2/9: Adults can do it by accessing the magnitude of the whole fractions. *Acta Psychologica.*

Meert, G., Grégoire, J., & of Psychology, M. (2010). Comparing the magnitude of two fractions with common components: Which representations are used by 10-and 12-year-olds? *Journal of Experimental Child Psychology.*

Michal, B., Hendler, T., Kahn, I., Dafna, B., & Grodzinsky, Y. (2003). The neural reality of syntactic transformations: evidence from functional magnetic resonance imaging. *Psychol Sci*, *14*(5), 433–440.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems.*

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K., Malave, V. L., a Mason, R., & Just, M. (2008). Predicting human brain activity associated with the meanings of nouns. *Science (New York, N.Y.)*, *320*(5880), 1191 1195. doi: 10.1126/science.1152876

Monti, M. M. (2011). Statistical analysis of fMRI Time-Series: a critical review of the GLM

approach. *Front Hum Neurosci*, *5*, 28. doi: 10.3389/fnhum.2011.00028

Monti, M. M. (2017). The role of language in structure-dependent cognition. *Neural mechanisms of language*.

Monti, M. M., & Osherson, D. N. (2012). Logic, language and the brain. *Brain research*, *1428*, 33 42. doi: 10.1016/j.brainres.2011.05.061

Monti, M. M., Osherson, D. N., Martinez, M. J., & Parsons, L. M. (2007). Functional neuroanatomy of deductive inference: a language-independent distributed network. *Neuroimage*, *37*(3), 1005-1016. doi: 10.1016/j.neuroimage.2007.04.069

Monti, M. M., Parsons, L. M., & Osherson, D. N. (2009). The boundaries of language and thought in deductive inference. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(30), 12554 9. doi: 10.1073/pnas.0902422106

Monti, M. M., Parsons, L. M., & Osherson, D. N. (2012). Thought beyond language: Neural dissociation of algebra and natural language. *Psychological Science*, *23*(8), 914-922. doi: 10.1177/0956797612437427

Morcom, A. M., & Fletcher, P. C. (2007). Does the brain have a baseline? why we should be resisting a rest. *Neuroimage*, *37*(4), 1073-82. doi: 10.1016/j.neuroimage.2007.06.019

Moyer, R., & Landauer, K. (1967). Time required for judgements of numerical inequality. *Nature*.

Mumford, J. (2014). Considerations when using Single-Trial parameter estimates in representational similarity analyses. In J. Eberwine (Ed.), (p. 44-50). Society for Neuroscience.

Mumford, J., Turner, B., Ashby, G., & Poldrack, R. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, *59*(3), 2636–2643. doi: 10.1016/j.neuroimage.2011.08.076

Murphy, B., Talukdar, P., & Mitchell, T. (2012). Learning effective and interpretable semantic models using non-negative sparse embedding. *Proceedings of COLING 2012*.

Musso, M., Weiller, C., Horn, A., Glauche, V., Umarova, R., Hennig, J., … Rijntjes, M. (2015). A single Dual-Stream framework for syntactic computations in music and language. *NeuroImage*, *117*, 267-83. doi: 10.1016/j.neuroimage.2015.05.020

Naccache, L., & Dehaene. (2001). The priming method: imaging unconscious repetition priming reveals an abstract representation of number in the parietal lobes. *Cerebral cortex*.

Naselaris, T., & Kay, K. N. (2015). Resolving ambiguities of MVPA using explicit models of representation. *Trends Cogn Sci*, *19*(10), 551-554. doi: 10.1016/j.tics.2015.07.005

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, *56*(2), 400-410. doi: 10.1016/j.neuroimage.2010.07.073

Nieder, A., & Dehaene. (2009). Representation of number in the brain. *Annual review of neuroscience*.

Nili, H., Wingfield, C., Walther, A., Su, L., William, M., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS computational biology*, *10*(4), e1003553. doi: 10.1371/journal.pcbi.1003553

Nishimoto, S., & Nishida, S. (2016). Lining up brains via a common representational space. *Trends Cogn Sci*, *20*(8), 565-567. doi: 10.1016/j.tics.2016.06.001

Ogawa, S., Lee, T., Kay, A., & Tank, D. (1990). Funtional brain mapping by blood oxygenation level dependent contrast magnetic resonance imaging: A comparison of signal characters with a biophisical model. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, *87*, 9868 - 9872.

Opfer, J., & Siegler, R. (2012). The oxford handbook of thinking and reasoning. In (chap. Development of quantitative thinking). Oxford University Press.

Osherson, D. (1975). Logic and models of logical thinking. *Reasoning: Representation and process in children and adults*, 81–91.

Patel, A. (2003). Language, music, syntax and the brain. *Nat Neurosci*, *6*(7), 674-681. doi: 10.1038/nn1082

Patel, A. (2012). Language, music, and the brain: a resource-sharing framework. *Language and music as cognitive systems*.

Patel, A., Gibson, E., Ratner, J., Besson, M., & of cognitive, H. P. (1998). Processing syntactic relations in language and music: An event-related potential study. *Journal of cognitive ….*

Patel, A., Iversen, J., Wassenaar, M., & Aphasiology, H. P. (2008). Musical syntactic processing in agrammatic broca's aphasia. *Aphasiology*. doi: 10.1080/02687030701803804

Patenaude, B., Smith, S., Kennedy, D., & Jenkinson, M. (2011). A bayesian model of shape and appearance for subcortical brain. *Neuroimage*, *56*, 907 - 922.

Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B., & Gramfort, A. (2015). Data-driven HRF estimation for encoding and decoding models. *NeuroImage*, *104*, 209 220. doi: 10.1016/j.neuroimage.2014.09.060

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Front Neuroinform*, *2*, 10. doi: 10.3389/neuro.11.010.2008

Pereira, F., Botvinick, M., & Detre, G. (2013). Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artif Intell*, *194*, 240-252. doi: 10.1016/j.artint.2012.06.005

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., … Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nat Commun*, *9*(1), 963. doi: 10.1038/s41467-018-03068-4

Peretz, I. (1993). Auditory atonalia for melodies. *Cognitive Neuropsychology*. doi: 10.1080/02643299308253455

Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature neuroscience*. doi: 10.1038/nn1083

Peretz, I., Kolinsky, R., Tramo, M., Labrecque, R., Hublet, C., Demeurisse, G., & Belleville, S. (1994). Functional dissociations following bilateral lesions of auditory cortex. *Brain*. doi: 10.1093/brain/117.6.1283

Peretz, I., Vuvan, D., Lagrois, M.-É. É., & Armony, J. L. (2015). Neural overlap in processing music and speech. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, *370*(1664), 20140090. doi: 10.1098/rstb.2014.0090

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron*, *53*(2), 293-305. doi: 10.1016/j.neuron.2006.11.022

Piccirilli, M., Sciarma, T., & Luzzi, S. (2000). Modularity of music: evidence from a case of pure amusia. *Journal of Neurology*. doi: 10.1136/jnnp.69.4.541

Pinel, P., Dehaene, S., Riviere, D., & LeBihan. (2001). Modulation of parietal activation by semantic distance in a number comparison task. *Neuroimage*.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci. (Regul. Ed.)*, *10*(2), 59-63. doi: 10.1016/j.tics.2005.12.004

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., … Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.*, *18*(2), 115-126. doi: 10.1038/nrn.2016.167

Posner, M., & Deheane, S. (1994). Attentional networks. *Trends in Neurosciences*, *17*, 75 - 79.

Rapp, M., Bassok, M., DeWolf, M., & Holyoak, K. J. (2015). Modeling discrete and continuous entities with fractions and decimals. *J Exp Psychol Appl*, *21*(1), 47-56. doi: 10.1037/xap0000036

Rissman, J., Gazzaley, A., & Mark, D. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *NeuroImage*, *23*(2), 752 63. doi: 10.1016/j.neuroimage.2004.06.035

Rogalsky, C., Rong, F., Saberi, K., & Hickok, G. (2011). Functional anatomy of language and music perception: temporal and structural factors investigated using functional magnetic resonance imaging. *Journal of …*. doi: 10.1523/JNEUROSCI.4515-10.2011

Rowling, J. (1998). *Harry potter and the sorcerer's stone*. Arthur A. Levine Books.

Scheperjans, F., Hermann, K., Eickhoff, S. B., Amunts, K., Schleicher, A., & Zilles, K. (2008). Observer-independent cytoarchitectonic mapping of the human superior parietal cortex. *Cereb. Cortex*, *18*(4), 846-67. doi: 10.1093/cercor/bhm116

Schneider, M., & Siegler, R. S. (2010). Representations of the magnitudes of fractions. *J Exp Psychol Hum Percept Perform*, *36*(5), 1227-38. doi: 10.1037/a0018170

Seyed-Mahdi, K., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *Plos Comput Biol*, *10*(11), e1003915. doi: 10.1371/journal.pcbi.1003915

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cogn Psychol*, *62*(4), 273-96. doi: 10.1016/j.cogpsych.2011.03.001

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, E. J., H, J., … Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, *23*, S208 - S219.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, *44*(1), 83-98. doi: 10.1016/j.neuroimage.2008.03.061

Sprute, L., & Temple, E. (2011). Representations of fractions: Evidence for accessing the whole magnitude in adults. *Mind, Brain, and Education*. doi: 10.1111/j.1751-228X.2011.01109.x

Steinbeis, N., & Koelsch, S. (2008). Shared neural resources between music and language indicate semantic processing of musical tension-resolution patterns. *Cereb. Cortex*, *18*(5), 1169-78. doi: 10.1093/cercor/bhm149

Stephan, K., Harrison, L., Penny, W., & Friston, K. (2004). Biophysical models of fMRI responses. *Current Opinion in Neurobiology*, *14*(5), 629–635.

Stout, F., & Kalivas, J. (2007). Evaluation of multivariate calibration using a tikhonov regularization approach and the generalized Pair-Correlation method with nonlinear data. *Analytical letters*, *40*(6), 1227-1251. doi: 10.1080/00032710701298529

Stout, F., Kalivas, J. H., & Héberger, K. (2007). Wavelength selection for multivariate calibration using tikhonov regularization. *Applied spectroscopy*, *61*(1), 85-95. doi: 10.1366/000370207779701479

Tanji, J., & Shima, K. (1994). Role for supplementary motor area cells in planning several movements ahead. *Nature*. doi: 10.1038/371413a0

Tanji, J., Shima, K., & Mushiake, H. (2007). Concept-based behavioral planning and the lateral prefrontal cortex. *Trends in cognitive sciences*.

Tettamanti, M., & Weniger, D. (2006). Broca's area: a supramodal hierarchical processor? *Cortex*, *42*(4), 491–494.

Thurman, S. M., van Boxtel, J. J., Monti, M. M., Chiang, J. N., & Lu, H. (2016). Neural adaptation in pSTS correlates with perceptual aftereffects to biological motion and with autistic traits. *Neuroimage*, *136*, 149-61. doi: 10.1016/j.neuroimage.2016.05.015

Tikhonov, A. (1943). On the stability of inverse problems. *Soviet Mathematics*, *39*(5), 195-198.

Tikhonov, A. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics*, *4*, 1035-1038.

Tillmann, B., Janata, P., & Research, B. J. (2003). Activation of the inferior frontal cortex in musical priming. *Cognitive Brain Research*.

Todd, M., Nystrom, L., & Cohen, J. (2013). Confounds in multivariate pattern analysis: theory and rule representation case study. *Neuroimage*.

Turner, B. O., Mumford, J. A., Poldrack, R. A., & Ashby, F. (2012). Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. *Neuroimage*, *62*(3), 1429-38. doi: 10.1016/j.neuroimage.2012.05.057

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., … Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, *15*(1), 273-89. doi: 10.1006/nimg.2001.0978

Varley, R. A., Klessinger, N. J., Romanowski, C. A., & Siegal, M. (2005). Agrammatic but numerate. *Proc Natl Acad Sci U S A*, *102*(9), 3519-3524. doi: 10.1073/pnas.0407470102

Vartanian, O. (2012). Dissociable neural systems for analogy and metaphor: Implications for the neuroscience of creativity. *British journal of psychology*, *103*(3), 302-316. doi: 10.1111/j.2044-8295.2011.02073.x

Vendetti, M. S., & Bunge, S. A. (2014). Evolutionary and developmental changes in the lateral frontoparietal network: a little goes a long way for higher-level cognition. *Neuron*, *84*(5), 906-17. doi: 10.1016/j.neuron.2014.09.035

Vogel, C. (2002). *Computational methods for inverse problems*. Society for Industrial and Applied Mathematics.

Waechter, R. L., Goel, V., Raymont, V., Kruger, F., & Grafman, J. (2013). Transitive inference reasoning is impaired by focal lesions in parietal cortex rather than rostrolateral prefrontal cortex. *Neuropsychologia*, *51*(3), 464-71. doi: 10.1016/j.neuropsychologia.2012.11.026

Waldorp, L. (2009). Robust and unbiased variance of GLM coefficients for misspecified autocorrelation and hemodynamic response models in fMRI. *International Journal of Biomedical Imaging*, *2009*(723912), 1–11.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016).

Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage*, *137*, 188-200. doi: 10.1016/j.neuroimage.2015.12.012

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE*, *9*(11), e112575. doi: 10.1371/journal.pone.0112575

Weisskoff, R., Baker, J., Belliveau, J., Davis, T., Kwong, K., Cohen, M., & Rosen, B. (1993). Power spectrum analysis of functionally-weighted MR data: What's in the noise? *Proceedings of the Society of Magnetic Resonance Medicine*, *1*, 7.

Wendelken, C. (2015). Meta-analysis: how does posterior parietal cortex contribute to reasoning? *Front Hum Neurosci*, *8*, 1042. doi: 10.3389/fnhum.2014.01042

Wendelken, C., Ferrer, E., Ghetti, S., Bailey, S. K., Cutting, L., & Bunge, S. A. (2017). Frontoparietal structural connectivity in childhood predicts development of functional connectivity and reasoning ability: A Large-Scale longitudinal investigation. *Journal of Neuroscience*, *37*(35), 8549-8558. doi: 10.1523/JNEUROSCI.3726-16.2017

Wendelken, C., Ferrer, E., Whitaker, K. J., & Bunge, S. A. (2016). Fronto-Parietal network reconfiguration supports the development of reasoning ability. *Cereb Cortex*, *26*(5), 2178-2190. doi: 10.1093/cercor/bhv050

Wertheim, J., & Ragni, M. (2018). The neural correlates of relational reasoning: A meta-analysis of 47 functional magnetic resonance imaging studies. *J Cogn Neurosci*, 1-15.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage*, *92*, 381-97. doi: 10.1016/j.neuroimage.2014.01.060

Woolrich, M., Ripley, B., Brady, J., & Smith, S. (2001). Temporal autocorrelation in

univariate linear modelling of fMRI data. *Neuroimage*, *14*, 1370 - 1386.