# Measuring Abstract Concepts using fMRI and the Implicit Semantic Accessibility Procedure: Implication for Measuring Representational Structures across Social Groups.

Roberto J. Vargas

Department of Psychology

Carnegie Mellon University

Pittsburgh, PA USA

June, 2022

**Committee:**

**Dr. Timothy Verstynen**

Associate Professor of Psychology and Neuroscience Institute, Carnegie Mellon University

**Dr. Sylvia Perry**

Associate Professor of Psychology, Northwestern University

**Dr. Kevin Jarbo**

Assistant Professor of Social and Decision Sciences, Carnegie Mellon University

**Dr. Leila Wehbe**

Assistant Professor of Machine Learning and Neuroscience Institute, Carnegie Mellon University

**Dr. Marcel Just (former advisor)**

D.O. Hebb Professor of Psychology, Carnegie Mellon University

## Abstract

Concepts are internal representations of characteristics of the external world with an ontological structure that associates related concepts together, like an apple tastes sweet and can sometimes be red. Concepts can be associated to form new ideas and concepts, can instantiate our lived experiences into language, and can be used to communicate our emotions and feelings towards people or institutions.

Although there has been extensive neuroscientific and behavioral research on the semantic properties of concrete concepts there has been far less work examining the underlying semantic structure of abstract concepts. Abstract concepts are not immediately experienced by taste, smell, hearing, touch, sight, via proprioception (awareness of our body's position in space or relative to an object), or via our vestibular system (information to maintain our body's orientation in space). These aperceptual abstract concepts are important in that they represent key domains of human knowledge, critical for many complex human decisions and allowing us to form new ideas based on old ideas. Neuroscientific research on the semantic structure of abstract concepts is sparse and the degree of homogeneity of meaning of abstract concepts across individuals remains a question. Moreover, the inverse question of whether differences in meaning for the same concept (abstract or concrete) can be measured still remains. Across people that share a common language there may be variation in concept associations driven by differences in experiences of those concepts.

This dissertation outlines three projects: First, fMRI-measured activation patterns were obtained during a thinking task for 28 abstract concepts (spanning 7 categories) from native English speakers. An examination of the underlying semantic structure revealed 3 dimensions underlying the 28 concepts, suggesting a brain-based ontology for this set of abstract concepts. The 3 dimensions corresponded to 1) the degree a concept was Verbally Represented; 2) whether a concept was External (or Internal) to the individual, and 3) whether the concept contained Social Content. Second, the degree of commonality of these dimensions across fMRI-measured activation patterns for the same 28 abstract concepts were compared across native Mandarin and English speakers. The semantic dimensions identified in the first project were replicated across languages; however, despite comparing different languages, there is no reason to hypothesize differences in concept meaning. Third, differences in behaviorally measured concept representations were assessed by measuring socioenvironmental concepts (e.g., *healthcare, police*) that have been shown to reflect racial disparities between White and Black Americans. Moreover, a novel implicit measure of semantic association, the Implicit Semantic Association Procedure, is proposed, and evaluated against gold-standard measures, as a tool for implicitly measuring concept semantic associations. Collectively this work measures the underlying structure of abstract concepts in the brain, the generalizability of these semantic structures across languages, and takes initial steps for measuring whether differences in experience leads to differences in concept representations.

## Acknowledgment

I initially want to thank Tim Verstynen. You have always been a role model to aspire to. You have taught me what it means to be a supportive and encouraging mentor and how to conduct science and research with integrity. You have most importantly taught me that academics can live (semi)normal lives too.

I would also like to thank my committee for providing insight and guidance on this project. Kevin Jarbo, particularly, has showed me what it means to strive for making academia a more inclusive place and what it means to leave a place better than how you found it.

To Marcel Just, for taking a chance on me and giving me the opportunity to achieve beyond my means. Also, Rob Mason, Vlad Cherkassy, and Tim Keller for having the patience to show a floundering research assistant how to navigate research at an elite institution.

To Amelia Van Howe, for being an incredibly supportive partner. You are color in an otherwise grey world.

To Cassie Eng, for being the best friend anyone could ask for. You have taught me that change can be more than just an idea. Also, that change is only a carefully worded email away. I would be a fraction of an academic if it wasn't for you.

Most of all I want to thank all the mentors who have taken a chance on me throughout my academic journey and making me feel like the luckiest academic on earth.

## Table of Contents

## Overview

Concepts are internal representations of characteristics of the external world with an ontological structure that associates related concepts together, like an apple tastes sweet and can sometimes be red. Concepts can be associated to form new ideas and concepts, can instantiate our lived experiences into language, and can be used to communicate our emotions and feelings towards people or institutions. Concepts also give us the means to communicate aperceptual information, also referred to as abstract concepts -- not immediately experienced by taste, smell, hearing, touch, sight, via proprioception (awareness of our body's position in space or relative to an object), or via our vestibular system (information to maintain our body's orientation in space). These aperceptual abstract concepts are important in that they are at the core of every domain of study scaffold the progress of human knowledge allowing us to form new ideas based on old ideas. Despite this importance there has been considerable research linking representations of object concepts and other perceptually grounded representations to the brain (Grill-Spector and Malach, 2004; Martin, 2007), our understanding of abstract concepts such as *gravity*, *spirituality*, and *prejudice* is far less understood. Neuroscientific research on the semantic structure of abstract concepts is sparse and the degree of homogeneity of meaning of abstract concepts across individuals remains a

question. Moreover, the inverse question of whether differences in meaning for the same concept (abstract or concrete) can be measured still remains. Across people that share a common language there may be variation in concept associations driven by differences in experiences of those concepts.

Chapter 1 consists of a published book chapter describing prevailing theories on the neuroscience of concepts, the literature on methodological approaches to understanding the link between concepts and the brain, and the current body of neurosemantic literature on concrete and hybrid concepts –concepts that are not concrete nor purely abstract (Vargas and Just, 2021).

Chapter 2 consists of a published article of one of the first MVPA examination of abstract concepts (Vargas and Just, 2020). In this study, fMRI-measures activation patters are obtained during a thinking task for 28 abstract concepts (spanning 7 categories) from a set of native English speakers. Using modern MVPA approaches, this study identified 3 semantic dimensions underlying the neural activation patterns of these 28 abstract concepts. The 3 dimensions corresponded to 1) the degree a concept was Verbally Represented; 2) whether a concept was External (or Internal) to the individual, and 3) whether the concept contained Social Content.

Chapter 3 is a published article which assesses whether the semantic dimensions identified in the study described in Chapter 2 are common across people of differing cultures and languages by examining the fMRI-measured neural activation patterns of the same 28 abstract concepts across American native English speakers and Chinese native Mandarin speakers (Vargas and Just, 2022). MVPA analyses revealed the same set of semantic dimensions underlying the set of abstract concepts including an additional dimension related to the concept's degree of reliance on rule-driven principles. Chapter 2 and 3 identified a common underlying semantic structure that spans languages and cultures but there was no basis to expect differences in the neural patters of the proposed set of 28 concepts besides polysemous meaning.

Chapter 4 aims extend upon Chapter 3 by describing a series of behavioral studies aimed at measuring differences in concept meaning (rather than commonalities) driven by differing experiences. Racial disparities between Black and White Americans have been well documents and empirically studied across numerous domains and institutions of society. This chapter contextualizes the Bias of Crowds model, which asserts that concept associations reflect environmental factors (broadly described as context and situations)(Payne et al., 2017). Contexts can be as proximal as a specific personal encounter or as distal as the systemic and cultural norms of the nation the individual resides in. Contexts can also be as isolated as a one-time incident or as

chronic as repeated exposure to associations such as media exposure. To measure differences in concept association, two gold-standard measures of concept geometry (Spatial Arrangement Method (SpAM) and the Pairwise Rating Method (PRaM)) will be utilized in addition to a novel proposed implicit measure of concept geometry, the Implicit Semantic Association Procedure (ISAP). The first study of this chapter assesses the reliability and validity of these three metrics using a set of simple object concepts and attributes with clearly defined associations. The second half of this chapter utilizes the measures of semantic association to assesses whether racial disparities between Black and White Americans are reflected in the association of concepts related to institutions within our American society (e.g., *healthcare, police*).

## Chapter 1. Approaches for understanding and analyzing the neural representations of concept

Although the study of concept knowledge has long been of interest in psychology and philosophy, it is only in the past two decades that it has been possible to characterize the neural implementation of concept knowledge. With the use of neuroimaging technology, it has become possible to ask previously unanswerable questions about the representation of concepts, such as the semantic composition of a concept in its brain representation. In particular, it has become possible to uncover some of the fundamental dimensions of representation that characterize several important domains of concepts.

Much of the recent research has been done with fMRI to predict and localize various concept representations and discover the semantic properties that underlie them. Commonly used experimental designs in this research area present single words or pictures of objects, measure the resulting activation pattern in multiple brain locations, and develop a mapping between the topographically distributed activation pattern and the semantic representation of the concept. The primary research topics concerning concept representations pertain to three issues: the composition of concept representations, the neurally defined underlying semantic dimensions; and the relation between neuroimaging findings and cognitive and psycholinguistic findings. It is these

types of relationships between cortical function and meaning representation that allow us to understand more about both the way knowledge is organized in the human brain and the functional role that various brain systems play in representing the knowledge.

Concepts are often qualitatively different from one another with regard to their perceptual grounding. As a result, one area of research has largely focused on the neural representations of concrete object concepts. However, as imaging technology and analytic techniques continue to improve, the neural representations of seemingly ethereal, abstract concepts such as ethics and truth have recently become a topic of increasing interest. In addition to the interest in such highly abstract concepts, recent research has also investigated hybrids between concrete and abstract concepts such as emotions, physics concepts, and social concepts. These hybrid concepts are not directly perceptually grounded but they can nevertheless be experienced. This chapter provides an overview of contemporary neuroimaging research examining the neural instantiation of concrete concepts, abstract concepts, and concepts that fall somewhere in between, which we call hybrid concepts.

*Hub-and-spoke model of feature integration in concept representations*

Any individual concept representation is thought to be composed of a network of semantic features (Collins and Loftus, 1975). Connections to more similar (closer) semantic representations are more likely and easier to come to mind than more distal ones. The anterior temporal lobe (ATL), sometimes referred to as the convergence zone or hub, has been credited with incorporating individual semantic features of concepts (the spokes, in this analogy) into an integrated representation of that concept (Meyer and Damasio, 2009). Recent fMRI research suggests that this integration of semantic features in the brain is localized to the ATL. One study showed that by combining color-related activation coded in right V4 and shape-related activation coded in the lateral occipital cortex (LOC) allowed visual objects to be distinguished in the ATL (Coutanche and Thompson-Schill, 2015). Although the ATL has also been shown to activate for abstract concepts (Hoffman 2016), a study similar to Coutanche and Thompson-Schill (2015) has yet to be conducted showing that individual abstract concepts can be decoded from ATL based on their composite semantic features. In sum, the ATL is thought to act as a cognitive mechanism that integrates perceptual and verbal (i.e., concrete and abstract) information comprising the representation of a concept (Lambon Ralph, 2014).

*General Semantic Network*

A prevailing neuroscientific framework for explaining concept and semantic processing is the General Semantic Network theory (Binder et al., 2009; Liuzzi et al., 2020). The authors computed a meta-analysis of 120 articles examining semantic and concept processing totaling 187 univariate activation maps. The activation maps were restricted to those which directly measured some form of semantic information. The 3 most common semantic maps included: Word versus pseudoword tasks that required participants to access word meaning; Semantic decision tasks with word stimuli; and high versus low meaningfulness tasks (e.g., names of animals versus tones; Binder et al., 1999). The meta-analysis converged upon a network of regions that could be attributable to semantic processing regardless of specific semantic categorization or task structure.

The result of this meta-analysis describes a set of brain locations responsible for the processing of semantic information (See Figure 1.1). This network of regions includes: 1) inferior parietal lobe (AG and portions of SMG), 2) lateral temporal cortex (MTG and portions of ITG), 3) ventral temporal cortex (mid-fusiform and adjacent parahippocampal gyrus), 4) DMPFC, 5) IFG, 6) ventromedial prefrontal cortex (VMPFC), and 7) posterior cingulate gyrus (Binder et al., 2009).

This framework provides a convenient vessel for modern multivariate pattern analytic (MVPA) approaches for understanding semantic structures underlying

qualitatively distinct concepts. Using this framework, semantic dimension identified for individual sets of concepts (e.g., *shelter* dimension, *frequency* dimension, *valance* dimension; Just et al.,2010) can be thought as being distributed sub-systems which serve as organization principles for processing specific types of information subsumed within this network.

**Figure 1.1.** Summary diagram of the regions comprising the general semantic network. This figure was adapted from Figure 7 as published in Binder et al., 2009.

### *Neuroscientific methods for understanding concepts*

*Univariate-based analyses*

The initial approach of task-related fMRI imaging was to measure the difference in activation for a class of stimuli (such as a semantic category, like houses) relative to a "rest" condition. At each 3-dimensional volume element in the brain (a voxel), a general linear regression model (GLM) is fit to relate the occurrence of the stimuli to the increase in activation relative to the rest condition. The result is a beta weight whose magnitude reflects the degree of condition-relevant activation in each voxel. This approach proves useful for investigating the involvement of cortical regions whose activation systematically increases or decreases relative to rest for a specific mental activity. However, with this voxel-wise univariate approach, complex relations between the activation in different brain regions within a network are often not apparent (Kriegeskorte, et al., 2006; Mur, et al., 2009). Moreover, treating each voxel independently of the others misses the fact that the activation pattern corresponding to a concept consists of a set of co-activating voxels that may or may not be proximal to each other. Nevertheless, the univariate approach was successful in identifying which brain regions activated in response to a given class of concepts.

*Multivariate pattern analysis (MVPA)*

The advent of higher-resolution imaging analyses aided in shifting the research focus from identifying the cortical regions involved in the representation of concepts to focusing on the coordinated activation across a network of brain regions or subregions (Haxby et al., 2001; Haynes and Rees, 2006). Instead of assessing the activation evoked by a class of concepts in terms of individual voxels in various brain regions considered independently of each other, multivariate analyses treated the activating voxels in conjunction with each other, as multiple dependent variables. Multivariate pattern analysis (MVPA) is graphically illustrated in Figure 1.2. MVPA refers to a family of analyses designed to take into account the multivariate relationships among the voxels that represent various concepts. Some of the most common analyses for investigating concept representations include: 1. Representational Similarity Analysis (RSA), which enables *comparison* of the multivariate activation patterns of different concepts; 2. Factor Analysis or Principle Components Analysis (PCA) which enables discovery of the lower-dimensional structure of distributed patterns of activation; 3. Predictive Modeling, which enables assessment of various postulated interpretations of underlying semantic structures by predicting activation patterns of concepts; and 4. Encoding Models which enable quantitative assessment of various organizational structures hypothesized to drive the activation. These techniques tend to answer somewhat

different questions.

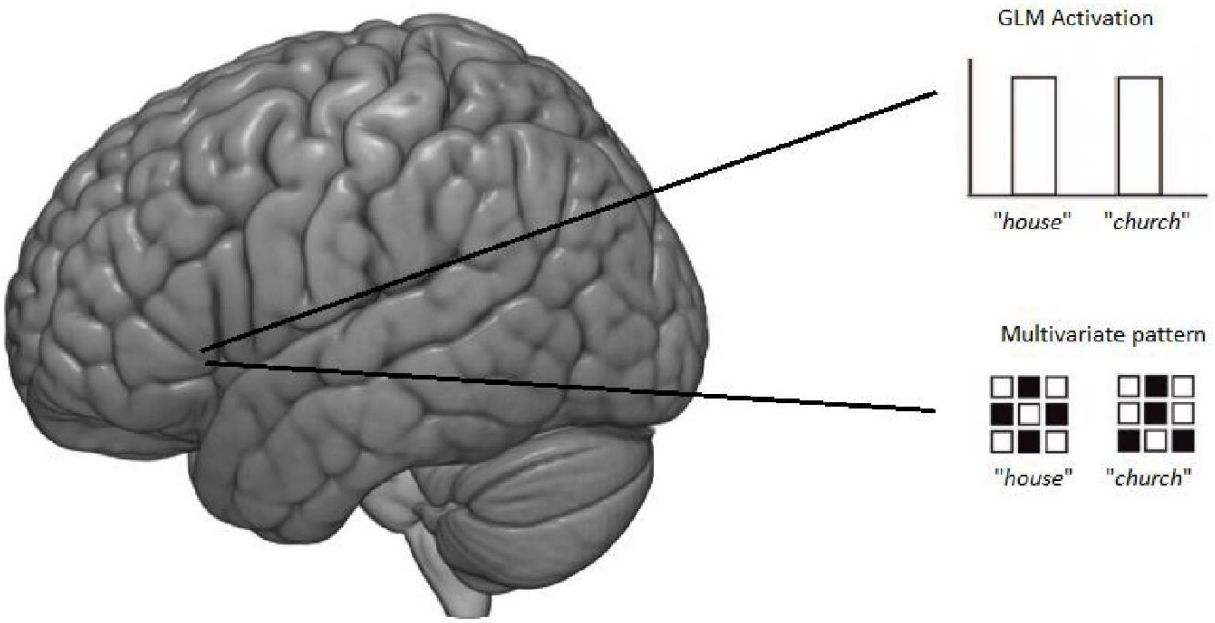


Figure 1.2 Conceptual schematic showing differences between GLM activation-based approaches and pattern-oriented MVPA, where the same number of voxels activate (shown as dark voxels) for two concepts but the spatial pattern of the activated voxels differs

Representational similarity analysis (RSA)

RSA is often used to measure the similarity (or dissimilarity) of representational structures of various individual concepts or categories of concepts. The representation of a concept or a category of concepts can be defined as the evoked activation levels of some set of voxels. These activation patterns can be computed with respect to all of the voxels in the whole cortex but are often restricted to the voxels in semantically relevant regions. The most common technique is to redefine the representation of a concept from being an activation pattern to a similarity pattern with respect to the other concepts in

the set (Kriegeskorte, et al., 2008a). For example, the neural representation of a concept like robin can be thought of in terms of its similarities to a set of other birds. This approach makes it possible to compare various brain subsystems in terms of the types of information they represent, and thus to characterize the processing characteristics of each subsystem. For example, RSA has been used to demonstrate the similarities in the visuospatial subsystems of humans and monkeys in the representations of visually depicted objects (Kriegeskorte et al., 2008b). The strength of this approach is its higher level of abstraction of the neural representation of concepts, representing them in terms of their relations (similarities) to other concepts. The cost of this approach is its limited focus on the representation of the properties of individual concepts.

Extracting dimensions of semantics (Factor Analysis / PCA)

Factor analysis and PCA are used to extract neurally meaningful dimensions from high-dimensional activation patterns. "Neurally meaningful" refers to a subset of concepts systematically evoking activation from a subset of relevant voxels. For example, concrete objects that entail interaction with parts of the human body (such as hand tools) evoke activation in motor and pre-motor areas, such that a neural dimension of body-object interaction emerges (Just et al., 2010). This approach focuses

on dimensions that are shared by some concepts and de-emphasizes the differences among the concepts that share the dimension. The regions corresponding to the dimension can be localized to particular brain areas (by noting the factor loadings of various clusters of voxels).

After the dimension reduction procedure finds a dimension and the items and voxels associated with it, the dimension requires interpretation. The source of the interpretation often comes from past knowledge of the functional roles of the regions involved and the nature of the items strongly associated with the dimension. For example, if hand tools obtain the highest factor scores on some factor, then that factor might plausibly be interpreted as a body-object interaction factor. (The items' factor scores indicate the strength of the association between the items and the factor). One approach to assessing an interpretation of a dimension (such as a body-object interaction dimension in the example above) is to first obtain ratings of the salience of the postulated dimension, say body-object interaction, to each of the items from an independent group of participants. For example, the raters may be asked to rate the degree to which a concept, such as pliers, is related to the hypothesized dimension body-object interaction (Just et al., 2010). Then the correlation between the behavioral ratings and the activation-derived factor scores of the items provides a measure of how well the interpretation of the dimension fits the activation data. This technique has been

used to extract and interpret semantically meaningful dimensions underlying the representations of both concrete nouns and abstract concepts (Just et al., 2010).

Predictive modeling

The goal of a predictive modeling procedure is to assess whether the activation pattern of a concept that was left out of the modeling can be predicted with reasonable accuracy, given some theoretical basis. The prediction process starts by generating a hypothesis about the underlying factor or dimension (which is based on how the items are ordered by their factor scores and on the locations of the voxels with high factor loadings). Then a linear regression model is used to define the mapping between the salience ratings of all but one item and the activation levels evoked by those items in factor-related locations (voxel clusters with high factor loadings in factor analyses that excluded the participant in question). The mapping is defined for all of the underlying factors. Then the activation prediction for the left-out item is generated by applying the mappings for all of the factors to the ratings of the left-out item. This process is repeated, each time leaving out a different item, generating an activation prediction for all of the items. Activation predictions for each concept can be made within each participant and then averaged over participants. The accuracy of the predictions provides converging evidence for the interpretation of the neurosemantic factors. Unlike

the correlational measure described above, this approach develops a mapping that is generative or predictive, applying to items uninvolved in the modeling.

Hypothesis-driven encoding modeling

Encoding models provide another more general way to test whether a hypothesized semantic organization structure is capable of explaining the activation data for some set of concepts. A first step in the modeling is the specification of a theoretically plausible feature set that is hypothesized to account for the relationship between a stimulus set and the corresponding evoked activation patterns (Naselaris et al., 2011). For example, the co-occurrence of noun concepts with verbs in a large text corpus may account for the relationship between individual concepts and activation patterns for those concepts say in a regression model. The resulting beta-weights from the regression model quantify the degree to which each feature determines the relationship between the stimuli and neural activity (Mitchell et al., 2008). The ability of this mapping to generalize to novel concepts, either in activation space or in feature space, provides a quantitative assessment of the plausibility of the hypothesized relation. This approach is especially useful for representations that are less clearly

mapped in the brain, such as abstract concepts, enabling an evaluation of the neural plausibility of theories of abstract concept representation (Wang et al., 2018).

More recently, encoding models have been used with semantic vectors, a feature structure constructed by extracting information from the co-occurrence of words in a large text corpus, to serve as a basis for predictions of large-scale sets of concept representations (Pereira et al., 2018). Encoding models have also been used to measure the ability for theoretically derived semantic feature structures to explain neural activation data for sentences (Yang et al., 2017). Encoding models are a flexible tool that allow for the quantitative evaluation of the ability of theoretically motivated feature structures to account for brain activation patterns.

*Neurosemantic structure of concrete object representations*

Object concepts are the most perceptually driven of concept representations. Consequently, the neural representation of object concepts is well understood because the neural organization of low-level perceptual information is well understood (Grill-Spector and Malach, 2004; Martin, 2007). Haxby and colleagues (2001) showed that pictures of different objects could be related to each other based on their pattern of activation in the visuospatial pathway, specifically in the fusiform face area (FFA) and

parahippocampal place area (PPA). Patterns of activation in these regions were distinguishable in terms of the object categories being represented (i.e., faces versus houses). It seems clear that a substantial part of concrete object representations consists of the representation of their perceptual properties.

Moreover, it has been possible to determine the sequence in which various types of perceptual information becomes activated as the thought of a concrete object emerges. Recent MEG research has shown that the temporal trajectory of the neural activation for object representations starts with low-level visual properties such as image complexity which begins to be activated about 75 ms after stimulus onset in the early bilateral occipital cortex. Later, at 80-120 ms, information concerning more complex categorically defined shapes (e.g., has eyes, has 4 legs) begins to be activated along the left ventral temporal cortex and anterior temporal regions (Clarke et al., 2013). The early onset object representation suggests that coarse categorical distinctions between objects are rapidly represented along a left-hemispheric feed-forward neural pipeline. After this initial representation is generated, more complex semantic features take form through recurrent activation and the integration of more distributed cortical systems at 200-300 ms. This temporal trajectory from simple to complex information suggests a cumulating pipeline designed to construct meaning from distributed semantic features.

Beyond the understanding that concrete object representations are based in large part on the objects' perceptual properties, several interesting questions remain, such as how the differing perceptual properties of an object are integrated in the object representation and what semantic properties underlie the organization of the representations of differing objects.

Semantic dimensions of concrete concepts

Contemporary research into concrete object concepts has progressed beyond the focus on perceptual aspects of concept representations and begun to examine higher-level semantic properties of concrete concept representations. This approach generally utilizes dimension reduction techniques such as factor analysis, first on an individual participant level then at the group level, to investigate semantic dimensions that are present in the neural representations across individuals (Just et al., 2014). This dimension reduction approach applied to a set of activation patterns has the advantage of discovering neurally driven dimensions of meaning rather than imposing a previously hypothesized semantic organization.

Just and colleagues (2010) utilized this approach to uncover 3 semantic dimensions underlying the representation of 60 words referring to concrete nouns (e.g.,

hammer, apple). Specifically, they found that these 60 concrete concepts could be characterized by the way they relate to eating, manipulation (or body-object interaction), and shelter (or enclosure). Moreover, each of these dimensions was associated with a small set of cortical regions. The shelter dimension was associated with activation in regions of bilateral parahippocampal place area, bilateral precuneus, and left inferior frontal gyrus. The manipulation dimension was associated with activation in regions of left supramarginal gyrus and left pre- and post-central gyrus (the participants were right-handed). The eating dimension was associated with activation in regions of left inferior and middle frontal gyrus and left inferior temporal gyrus. These results indicate the beginnings of a biologically plausible basis set for concrete nouns and highlight semantic properties beyond a visuospatial domain.

Other research has sought to discover semantic dimensions of non-word or picture concept representations using a different approach. Principal Components Analysis (PCA) was applied to the activation evoked by 1800 object and action concepts shown in short movie clips (Nishimoto et al., 2011). This approach sub-divided the brain based on the similarities of the activation patterns among the concepts to their co-occurrence with a large text corpus. This technique was also applied to the activation patterns evoked by natural continuous speech (Huth et al., 2016). Both the video clip and the natural-speech studies related neural activation similarities to corpus co-

occurrence information to locate semantically consistent regions within the cerebral cortex based on domain-specific information. This parcellation approach associated the activation of various regions and semantic categories with individual concepts. The 12 interpretable semantic categories from the PCA were: mental (e.g., *asleep*); emotional (e.g., *despised*); social (e.g., child*);* communal (e.g., *schools*); professional (e.g., *meetings*); violent (e.g., *lethal*); temporal (e.g., *minute*); abstract (e.g., *natural*); locational (e.g., *stadium*); numeric (e.g., *four*); tactile (e.g., *fingers*); and visual (e.g., *yellow*). Aside from the format of stimulus presentation, the notable distinction between the dimension reduction approaches in Just et al. (2010) and Huth et al. (2016) was that Huth generated semantic dimensions based on the mapping between activation and co-occurrence while Just (2010) generated dimensions from the activation patterns. The exploration of the underlying dimensions of concrete concepts helps provide a basis for the semantic organization of perceptible concepts beyond basic visuospatial properties.

*Neurosemantic signatures of abstract concepts*

The representations of abstract concepts, such as ethics and law, are neurally and qualitatively distinct from those of concrete concepts. Abstract concepts, by definition, have no direct link to perception with the exception of some form of symbolic

representation (e.g., lady justice holding a scale to represent the concept of law or justice). The conventional view of abstractness portrays it as an absence of a perceptual basis, that is, the opposite of concreteness (Barsalou, 1999; Barsalou, 2003; Brysbaert et al., 2014; Wang et al., 2010). Although it is easy to define abstract concepts as those lacking concreteness, this definition does not describe the psychological or neurocognitive properties and mechanisms of abstract concepts.

Concrete and abstract concepts generally evoke different activation patterns, as a meta-analysis showed (Wang et al., 2010). This meta-analysis indicated that the two types of concepts differ in their activation in areas related to verbal processing, particularly the left inferior frontal gyrus (LIFG). Abstract concepts elicited greater activation than concrete concepts in such verbal processing areas. By contrast, concrete concepts elicited greater activation than abstract concepts in visuospatial processing (precuneus, posterior cingulate, and fusiform gyrus). This meta-analysis was limited to univariate comparisons of categories of concepts and did not have access to the activation patterns evoked by individual concepts. This limitation potentially overlooks nuanced distinctions in the representational structure. As described above, univariate contrasts potentially overlook critical relationships across neural states and neural regions (Mur et al., 2009). Through the use of MVPA techniques, more recent studies have begun to examine the underlying semantic structure of sets of abstract concepts.

The section below focuses on various imaging studies examining the neural activation patterns associated with abstract concepts and explores the possible semantic structures that are specific to abstract concepts.

As in the case of concrete concepts, the semantic dimensions underlying abstract concept categories can be identified from their activation patterns. One of the first attempts to decode the semantic content of abstract semantic information was conducted by Anderson and colleagues (2017). A set of individual concepts that belonged to various taxonomic categories (tools, locations, social roles, events, communications and attributes) were decoded from their activation patterns. Whether a concept belonged to 1 of 2 abstract semantic categories (i.e., *Law* or *Music*) was also decoded from the activation patterns of individual concepts. Although these abstract semantic categories could be decoded based on their activation patterns, the localization of this dissociation is unclear.

*Hybrid concepts: neither completely concrete nor completely abstract*

Hybrid concepts are concepts that can be experienced directly but require additional processing beyond the 5 basic perceptual faculties to evoke. These concepts do not neatly fit within the dichotomy of concrete versus abstract. For example, the

concept envy cannot be tasted, seen, heard, smelled, and touched but it inarguably can be experienced as an internal event which could have perceptual repercussions (e.g., feeling lethargic, crying). We propose that envy and other concepts referring to psychological states are hybrid.

The view of embodied cognition (Barsalou, 1999) expands upon the definition of perception beyond our 5 basic perceptual faculties to also include the experiences of proprioception and emotions. Hybrid concepts fall outside the strict realm of the sensory-perceptual but within the realm of psychological experience as described by embodiment theory (e.g., proprioception and emotion). Emotions, physics, and social concepts are not usually defined exclusively with respect to their concreteness/abstractness but serve as excellent exemplars of hybrid concepts in that they can be perceptually experienced without evoking any of our 5 senses directly. Additionally, the neural understanding of the semantic underpinning of hybrid concepts is not well understood. The following three sections describe research investigating the neurosemantic organization of hybrid concepts, specifically, the neurosemantic organization of emotions, physics concepts, and social concepts.

Neurosemantic dimensions of meaning underlying emotions concepts

Meta-analyses of activation contrasts investigating emotion concepts reveal six functional networks (Kober et al., 2008) including limbic regions (i.e., amygdala, hypothalamus, and thalamus), areas related to top-down executive control function (i.e., dorsal lateral prefrontal cortex), the processing of autobiographical information (i.e., posterior cingulate cortex) (Klasen et al., 2011), visual association regions, and subregions within the motor cortex (Phan et al., 2002). These networks suggest that emotion representations partially involve cognitive functions related to more complex perceptual functioning (i.e., motion and visual association). Furthermore, the involvement of regions related to top-down executive functioning and regions related to the processing of autobiographical information suggest that emotion concepts recruit cognitive faculties for not only basic perceptual representations but also for higher-ordered cognitive functions. Although these findings identify regions involved in emotion representation and processing, they do not provide insight into how different emotions are neurally distinguished.

Recent MVPA analyses examining the neural representations of emotion concepts have provided insight into the way the representations of different emotions are neurally organized. Kassam and colleagues (2013) examined the evoked neural activation patterns of 18 emotion concepts such as happiness, pride, envy and sadness. The participants in this study didn't just think about the meaning of a presented

emotion word: they tried to evoke the emotion in themselves at that moment. Factor analyses of the activation profiles for the 18 emotion concepts followed by a predictive model to validate the interpretations revealed three underlying dimensions of meaning. The underlying semantic dimensions organize these emotions concepts according to the valence of the emotion (positive or negative), its degree of arousal (fury versus annoyance) and degree of social involvement (i.e., whether another person is included in the representation, as is the case for envy but not necessarily so for sadness). Each of these dimensions of representation were found to correspond to activation distributed across several cortical regions. The brain locations associated with the valence of an emotion concept included the right medial prefrontal cortex, left hippocampus, right putamen, and the cerebellum. The brain locations associated with the arousal dimension included the right caudate and left anterior cingulum. Finally, brain locations associated with sociality included the bilateral cingulum and somatosensory regions. Both univariate and multivariate approaches provided neural evidence for the involvement of perceptual and higher-cognitive faculties. The multivariate analyses provided additional insight into the dimensions along which the individual emotion concepts are differentiated from each other. So even though emotions are very different from object concepts, the principles underlying their neural representations are rather similar to those of other types of concepts.

Neurosemantic dimensions of meaning underlying physics concepts

Research investigating the neural representation of physics concepts suggests their neural organization somewhat reflects the physical world they refer to, such as the movements or interactions of objects. Mason and Just (2016) investigated the neural activation patterns of 30 elementary physics concepts (e.g., acceleration, centripetal force, diffraction, light, refraction). Factor analyses of the activation patterns evoked by the 30 concepts revealed four underlying semantic dimensions. These dimensions were periodicity (typified by words such as wavelength, radio waves, frequency), causal-motion/visualization (e.g. centripetal force, torque, displacement), energy flow (electric field, light, direct, current, sound waves, and heat transfer), and algebraic/equation representation (velocity, acceleration, and heat transfer) which are associated with familiar equations.

The regions associated with each semantic dimension provide insight into the underlying cognitive role of the region. The periodicity dimension was associated with dorsal premotor cortex, somatosensory cortex, bilateral parietal regions, and the left intraparietal sulcus. These regions have been shown to activate for rhythmic finger tapping (Chen et al., 2006). The causal-motion/visualization dimension was associated

with the left intraparietal sulcus, left middle frontal gyrus, parahippocampus, and occipital-temporal-parietal junction. These regions have been shown to be involved in attributing causality to the interactions between objects and data (Fugelsang et al., 2005; Fugelsand and Dunbar, 2005). The algebraic/equations dimension includes the precuneus, left intraparietal sulcus, left inferior frontal gyrus, and occipital lobe. These regions have been implicated in the executive processing and integration of visuospatial and linguistic information in calculation (Benn, et al., 2012) and more general arithmetic processing. The regions associated with energy flow were middle temporal and inferior frontal regions. In the context of physics concepts, these regions are attributed with representing the visual information associated with abstract concepts (Mason and Just, 2016). Together, these results suggest that the neural representations of physics concepts, many of them developed only a few hundred years ago, draw on the human brain's ancient ability to perceive and represent physical objects and events.

Neurosemantic dimensions of meaning underlying social concepts

Research comparing the neural representation of social concepts between healthy controls and individuals with high-functioning autism has revealed three semantic dimensions involved in the neural representations of social interactions (Just et al., 2014). Participants in this study thought about the representations of 8 verbs describing

social interactions (compliment, insult, adore, hate, hug, kick, encourage, and humiliate) considered from the perspective of either the agent or recipient of the action. Factor analyses of neural activation profiled for these 16 concept-role combinations revealed semantic dimension associated with self-related cognition (hate in the agent role and humiliate in the recipient role), social valence (adore and compliment), and accessibility/familiarity relating to the ease or difficulty of semantic access.

The *self* dimension was associated with activation in posterior cingulate: an area commonly implicated in the processing of autobiographical information. The social valence factor included the caudate and putamen for both controls and individuals with autism. The accessibility/familiarity factor included regions that are part of the default mode network, particularly middle cingulate, right angular gyrus, and right superior medial frontal.

Because this study involved a comparison between young adult healthy controls and participants with high-functioning ASD, it provided an important glimpse into how a psychiatric or neurological condition can systematically alter the way a certain class of concepts is thought about. The use of fMRI neuroimaging allows the precise measurement of how a given concept is neurally represented, and specify precisely how a condition like ASD can alter the representation. The interesting finding was that the members of the two participant groups could be very accurately distinguished by their

neural representations of these social interaction concepts. More specifically, the ASD group showed a lack of a *self* dimension, showing little activation in the regions associated with the *self* dimension in the healthy control group. The findings suggest that when the ASD participants thought about a concept like hug, it involved very little thought of themselves. By contrast, the control group thought about themselves when thinking about what hug means. Thus the assessment of neural representations of various classes of concepts has the potential to identify the presence and the nature of concept alterations in psychiatric or neurological conditions. The neurosemantic architecture of hybrid concepts (as exemplified by emotions, physics, and social concepts) suggests these concepts relate us with the external world (e.g., causal-motion visualization dimension for physics concepts or self/other for social concepts). Moreover, the neural activation associated with magnitudes of perceptual experience are also captured by the neural representations (e.g., degree of arousal with emotion concepts). Taken together, these results suggest that hybrid concepts are composed, in part, of perceptual states that translate our perceptual world into various mental states.

*Relating neuroimaging findings and corpus co-occurrence measures*

One particular class of encoding models, as was previously discussed, attempts to relate neural representations to some well-defined feature set. Defining the meaning of a concept in some computationally tractable way has long been a challenge and it is relevant here because it has the potential to be systematically related to the neural representation of the concept. One of the early answers to this challenge suggested that concepts can be characterized in terms of the concepts with which they co-occur in some large text corpus (Landauer & Dumais, 1997). The lower dimensions (about 300) of a large co-occurrence matrix produce a semantic vector representation of the words in the corpus (Pennington et al., 2014; Deerwester et al., 1990). The method of deriving this lower dimensional feature space can vary, depending on the specific approach. The utility of semantic vector representations comes from their convenience in natural language processing applications. But can the semantic vector representation of a concept like apple be informative about the neural representation of apple?

The semantic vector representations can be used as the predictive basis of an encoding model. Predicted images can be generated from the learned mapping relating brain activation data from a matrix containing semantic vectors. This learned mapping can then be used to generate predicted brain images for concepts with no previously collected data (Mitchell et al., 2008). This approach provides the basis for generating a set of concept representations which can then be explored for its semantic properties

(Pereira et al., 2018). Moreover, it enables the study of many more concept representations than can easily be acquired in time and cost-limited fMRI studies. However, it is unclear whether encoding models based on semantic vector representations illuminate the difference between concrete and abstract concepts representations.

Co-occurrence structures have also been used to evaluate the neural instantiation of the associative theories of abstract concept representations. Wang and colleagues (2018) utilized RSA to compare the organizational structures of 360 abstract concept representations by examining the representational structure of fMRI activation patterns across the whole brain and concept co-occurrence properties in a large corpus. The goal was to show that each of these viable organization principles is instantiated uniquely within the brain.

Co-occurrence properties represent the theoretical view that abstract concepts are represented in terms of their association with other concepts. Their results showed that the relationship between co-occurrence representations and brain activity for 360 abstract concepts was largely left lateralized and seemed to uniquely activate areas traditionally associated with language processing such as left lateral temporal, inferior parietal, and inferior frontal regions.

*Conclusions*

The understanding of how concepts are represented in the human brain has advanced significantly based on innovations in imaging technology and multivariate machine learning techniques. One new insight concerns how human and self-centric concept representations are. No dictionary definition had specified how a hammer is to be wielded, and yet that is an important part of how it is neurally represented. Thus, part of the neural representation of a physical object specifies how our bodies interact with the object (Hauk and Pulvermuller, 2004; Just et al., 2010). Part of the neural representation of gossip specifies a social interaction. The concept of spirituality evokes self-reflection. Thus this insight is that many neural representations of concepts contain human-centric information in addition to semantic information.

A second insight concerns the dependence of abstract concepts on the verbal representations of other concepts. Representing the meaning of abstract concepts may require a greater integration of meaning across multiple other concept representations than is the case for concrete concepts. Abstract concepts evoke activation in cortical regions associated with language processing, particularly LIFG, which may reflect the neurocomputational demand for this increased integration of meaning.

A third insight is that the semantic components of a neural representation of a concept consist of the representations within various neural subsystems, such as the motor system, the social processing system, and the visual system. These neural subsystems constitute the neural indexing or organizational system.

A fourth insight concerns the remarkable degree of commonality of neural representations across people and languages. Although concept representations phenomenologically seem very individualized, the neural representations indicate very substantial commonality, while still leaving room for some individuality. The commonality probably arises from the commonality of human brain structures and their capabilities, and from commonalities in our environment. We all have a motor system for controlling our hands, and all apples have a similar shape, so our neural representations of holding an apple are similar.

A fifth insight is that the principles regarding the neural representations of physical objects extend without much modification to more concrete and hybrid concepts. Although it is easy to see why the concept of apple is similarly neurally represented in all of us, it is more surprising that an emotion like anger evokes a very similar activation pattern in all of us. Moreover, even abstract concepts like ethics have a systematic neural representation that is similar across people.

Although there is much more to human thought than the representation of concepts, these representations constitute an important set of building blocks from which thoughts are constructed. The neuroimaging of these concept representations reveals several of their important properties as well as hints as to how they might combine to form more complex thoughts.

# Chapter 2. Project 1: The Neural Representation of Abstract Concepts

Chapter 1 described the literature on the neuroscientific examination of concepts by briefly describing theoretical frameworks such as the hub-and-spoke model and the general semantic network model of how concepts are represented in the brain. Additionally, Chapter 1 described contemporary methodological approaches for understanding underlying semantic structure of concepts from brain activation using multivariate pattern analytic techniques. Lastly, Chapter 1 described how these methodologies allowed for the examination of concrete and hybrid concepts (e.g., physics and emotion concepts) in the brain. Chapter 2 describes a published empirical study examining the neural representation of abstract concepts using similar methodologies described in Chapter 1.

## Introduction

The human ability to formalize planetary orbit, argue what is ethical or just, or communicate about the feelings of others hinges on our ability to speak of concepts that do not explicitly take a physical form, or, abstract concepts. However, the neural characterization of abstract concepts such as ethics and justice remains relatively unexplained. A concept can be defined, in neural terms, as a systematic, distributed

pattern of activation across a network of cortical regions that occurs when a person thinks about that concept. Unlike concrete concepts, there are no explicit cortical systems or theories for explicitly measuring the embodied instantiation of abstract information. According to the now well-documented embodiment hypothesis, the representation of many concepts is rooted in how their referents are perceived and interacted with. This view has provided a valuable theoretical basis for understanding the neural instantiation of concrete concepts (Barsalou, 1999). However, there is much less clarity concerning the neural representation of abstract concepts (Binder et al., 2005).

A meta-analysis of functional magnetic resonance imaging (fMRI) studies examining concrete and abstract concepts revealed that areas responsible for language processing (namely, left inferior gyrus) reliably activate more for abstract concepts relative to concrete concepts (Wang et al., 2010). In addition, a number of studies have shown that several other cortical areas related to executive functioning, motion, and emotion processing were also involved in the processing of abstract concepts (Pecher et al., 2011; Vigliocco et al., 2014). These varied cortical activation findings suggest that abstract concept representations rely on the integration of multiple neural systems associated with a variety of cognitive functions.

The aperceptual nature of abstract concepts also raises the question of the commonality of their neural representations across individuals. Whereas the neural commonality of concrete concepts across people (Just et al., 2010) could be based on common perceptual properties, it is unclear what the common basis might be for more abstract concepts. The concept of justice, for example, is likely to be related to a wider variety of experiences than a concept such as apple, suggesting that the neural activation pattern associated with the concept could vary substantially across individuals. Previous research has shown that concrete concepts can be decoded across individuals from their neural signature; this commonality of representation can be characterized by lower-dimensional semantic primitives such as eating and shelter (Just et al., 2010; Coutanche & Thompson-Schill, 2015). This method of decoding concepts from neural signatures and characterizing their commonality across participants has been applied to perceptually less grounded categories of concepts such as physics concepts (Mason & Just, 2016) and emotion concepts (Kassam et al., 2013). However, although physics terms and emotions concepts are less concrete than apple or hammer, according to an embodied view of concept representation, they are still related to proprioception and emotional content. Thus, the commonality of the neural representation of abstract concepts across participants remains unknown.

The goal of the current study was to determine the neural and semantic ontology of individual abstract concepts. Although at least one previous study used multivariate pattern analytic techniques (MVPA) to decode taxonomic categories and domains of abstract concepts such law and music (Anderson et al., 2014), there has been no attempt to predict the neural representation of individual abstract concepts nor to uncover the semantic organization of abstract concepts in a neurally-based ontology. The current study assessed the neural activation patterns of 28 abstract concepts by applying MVPA including factor analysis to fMRI data. First, the identifiability and commonality of the concepts' neural signatures were assessed within and across participants using a pattern classifier. Second, a dimension-reduction technique (factor analysis) was used to derive a lower-dimensional semantic structure of the concept representations. These interpretations of the resulting semantic dimensions were then tested by obtaining independent ratings of the concepts along each of the dimensions as we interpreted them, and then using the ratings to predict the concepts' activation patterns. These findings provide a brain-based account of the way abstract concepts are neurally represented.

## Methods

*Participants*

Ten right-handed adults (7 Females; age range from 20 to 38, M=25.9) from the Carnegie Mellon community participated in a 30 min-scanning session. Informed consent was obtained from all 10 participants in accordance with the Carnegie Mellon Institutional Review Board. Data from 1 participant was excluded due to the participant falling asleep during the scan.

*Experimental Paradigm*

Stimuli were 28 words referring to abstract concepts distributed among 7 categories. Although the category labels were never mentioned nor presented to participants, they are listed here in parentheses for expository purposes, preceding the actual stimuli: (**mathematics**): *subtraction*, *equality*, *probability*, and *multiplication*; (**scientific**): *gravity*, *force*, *heat*, and *acceleration*; (**social**): *gossip*, *intimidation*, *forgiveness*, and *compliment*; (**emotion**): *happiness*, *sadness*, *anger*, and *pride*; (**law**): *contract*, *ethics*, *crime*, and *exoneration*; (**metaphysics**): *causality*, *consciousness*, *truth*, and *necessity*; (**religiosity**): *deity*, *spirituality*, *sacrilege*, and *faith*. Focusing on the neural representations of individual concepts provides a higher resolution of

semantic content than examination on a categorical level. The representations of individual concepts contain information about item-level elements of meaning rather than superordinate representational structures. The set of 28 stimuli was presented 6 times, to enable averaging out effects of noise in the fMRI signal and to provide separate datasets for training and testing the machine learning classifier in its cross-validation protocol. On each trial, participants were presented with the stimulus concept for 3 s, and were asked to think about the properties they associate with the given concept. Participants were instructed to think of the individual concept and the various components of its meaning, referring back to the properties of the concept they had generated. This instruction has previously been used to enable participants to evoke semantically rich representations of concepts that are consistent across multiple presentations (Just et al., 2010, 2017; Mason & Just, 2016; Bauer & Just, 2017).

Following this 3 s period, participants were instructed to clear their mind over the course of 7 s while watching a blue ellipse shrink to nonexistence, to allow the hemodynamic response to approach baseline before the next concept appeared. A shrinking ellipse was presented during the inter-stimulus interval to provide a fixation target and to convey the progress through the 7 s interval. There was a total of 6 presentation blocks of the same 28 stimulus concepts (using different random permutation orders in the different presentations) in the scanning session, distributed

between 3 runs (2 blocks per run) to allow participants a brief rest between runs. A 17-second "*X*" was presented at the beginning of each block (2 per run) to use as a baseline measure of neural activity.

Prior to the scan, participants were instructed to write down 3 properties for each of the 28 abstract concepts. Possible properties were synonyms, definitions, or experiences associated with the concept intended to guide participants to mentally evoke a consistent representation for each concept. Participants were instructed to write properties that came to mind quickly and naturally. Participants briefly practiced the experimental paradigm in a mock MRI scanner while receiving head-motion feedback to minimize movement.

*fMRI Parameterization and Image Processing*

Functional images were acquired on a Siemens Verio 3.0 T scanner and a 32-channel phased-array head coil (Siemens Medical Solutions, Erlangen, Germany) at the Scientific Imaging and Brain Research facility at Carnegie Mellon. Scans were acquired using a gradient-echo echo-planar imagining pulse sequence (TR=1000 ms, TE=25 ms, and a 60◦ f lip angle); each volume contained 20 5-mm thick AC-PC aligned slices (1-mm gap between slices). The acquisition matrix was 64 × 64 with 3.125 × 3.125 × 5-mm

voxels. SPM8 (http://www.fil.ion.ucl.ac.uk/spm/) was used to correct for head motion and normalize to the Montreal Neurological Institute template (Collins et al., 1994). The percent signal change (PSC) relative to the fixation condition was computed at each gray matter voxel for each stimulus presentation (the PSC data was converted to z-scores). To isolate the neural instantiation of concept representations, voxel activation levels were averaged over the four brain images acquired within a 4 s window (at a TR of 1000) offset 5 s from the stimulus onset (i.e., images 5 to 8). Mean PSCs were normalized across voxels for each trial (MPSC). Previous studies have reported that the use of these four images yields the highest classification accuracies obtained by a classifier that attempts to relate the activation pattern to the concept (Just et al., 2010; Mason & Just, 2016; Bauer & Just, 2017). Additionally, using these four images allows for the comparison with previously collected concept-level fMRI data.

*Voxel Stability*

The analysis focused on the most stable voxels, those whose activation levels were systematically modulated by the set of 28 abstract concepts each time the set was presented. Voxel stability is a criterion for feature selection that selects voxels in the training set that respond consistently across repetitions of the concepts across blocks. It

has been established as a method of feature selection for discriminating concept representations (Just et al., 2010, 2017; Kassam et al., 2013; Wang et al., 2013; Mason & Just, 2016; Bauer & Just, 2017; Yang et al., 2017). A voxel's stability was computed as the mean pairwise correlation of its 28 MPSC activation levels (for the 28 abstract concepts) across all pairwise combinations of the presentations blocks in the training data. Thus, a voxel with high stability is one that has a stable tuning curve over the set of stimuli. Stable voxels were used as features in classification and factor analyses. The stable voxels selected in the training data for classification are then used to select the voxels in the test set. The 120 most stable voxels in the whole brain were used as features for classification. This approximate number of voxels has been shown to reliably capture meaningful information in the neural representation of individual concepts (Just et al., 2010; Mason & Just, 2016). To ensure the analysis was not particularly sensitive to variations in the number of features, the classification analysis was repeated varying the number of stable voxels used from 20 to 10 000 (in 20 voxel increments); the peak classification accuracy occurred between 120 and 180 stable voxels. The mean classification accuracy gradually decreased with the inclusion of additional stable voxels beyond 180. To be consistent with previous studies, 120 stable voxels were chosen to be used as features.

*Discriminative Classification: Within-Participant*

Gaussian Naïve Bayes (GNB) classifier was trained to decode the 28 concepts in each participant's data. The classifier was trained on the activation data from 4 of the 6 presentations and was tested on the mean of the 2 left-out images. This cross-validation procedure was followed in 15 (6 choose 2) folds. The features used by the algorithm consisted of the activation levels of the 120 most stable voxels in the training set from anywhere in the whole brain. The classifier's normalized rank accuracy was used to assess decoding accuracy (i.e., the mean over folds of the normalized rank of the correct response in a probability-ranked list of all 28 alternatives, where the chance level is 0.5). Above-chance performance at $p < 0.001$ was achieved for concept-level predictions for all participants, as determined using a 10,000-iteration permutation test on each participant separately (mean cutoff for $p < 0.001 = 0.60$; SD= 0.004).

*Discriminative Classification: Between-Participant*

A GNB classifier was trained on the neural signatures from 8 of the 9 participants and tested on the left-out participant's data. The alignment across participants was accomplished by selecting the voxels with the highest stability across participants (i.e., having a similar pattern of activation responses to the 28 stimuli). To compute the

cross-participant stability in the between-participant classification, the MPSC data were first averaged across presentations for each participant and then the mean pairwise correlation of a voxel's 28 MPSC activation levels (for the 28 abstract concepts) was computed between all pairs of the 8 participants in the training data. The 120 most stable voxels (i.e., those with the highest mean pairwise correlations) from the whole brain across the 8 participants were selected as features for the training set. Predictions were cross-validated across participants and the mean rank accuracy was computed across the resulting 9 folds. Above-chance performance at $p < 0.01$ is 0.57 for concept-level predictions as determined using a 10,000-iteration permutation test.

*Factor Analysis Procedure*

To explore the semantic structure underlying the representations of the 28 abstract concepts, a two-level factor analysis was computed; a factor analysis was first applied to the data of individual participants while the second factor analysis used the factor scores from the first level as input (using a procedure described in detail in Just et al., 2014). This procedure was implemented using a principal factor analytic algorithm, including varimax rotation, in MATLAB (Version 6.5; The MathWorks, Natick, MA).

The data from all 9 participants were analyzed to determine whether interpretable factors could be extracted. Stability was averaged across the 9 participants for each voxel (voxels with negative stability were set to 0). The locations of the 800 most stable voxels were first used to indicate the major participating cortical regions [as defined using Automated Anatomical Labeling (AAL; Tzourio-Mazoyer et al., 2002)] to be included in the factor analysis. Then, the input to the first-level factor analysis (performed within each participant) consisted of the mean activation levels of the most stable voxels for each of the concepts in each of the contributing AAL regions. The total number of voxels used in this factor analysis was 410, similar to the number used in previous studies (Kassam et al., 2013), with the number per AAL-defined ROI based on the numerosity of the ROI's stable voxels: 40 voxels from left inferior frontal gyrus (LIFG); 30 voxels from left posterior cingulate cortex; 60 voxels from frontal cortex bilaterally; 60 voxels from occipital cortex bilaterally; 60 voxels from temporal cortex bilaterally; and 160 voxels from parietal cortex bilaterally. This first-level factor analysis was run on all 9 participants individually, extracting 7 factors for each subject, resulting in a total of 63 vectors of factor scores. The number of factors to be extracted was informed by previous studies (Mason & Just, 2016); modifications from the initial parameterization resulted in only minor differences in results.

The goal of the first-level factor analysis, applied to individual participants, was to partition the set of input stable voxels into subsets that each systematically but differentially responded to the abstract concepts, specifying 7 factors. This analysis produced factor scores for the 28 concepts, for each of the 7 factors, for each of the 9 participants. Each of the 9 participants' 7 sets of factor scores were concatenated and used as input into the second, group-level factor analysis (a total of 63 sets of 28 factor scores) to further reduce dimensionality to 5 dimensions and to seek consistency across participants. A voxel was determined to belong to a factor if its factor loading exceeded a cutoff 0.4 (a typical value for a factor loading threshold): this threshold was also informed by previous work using this procedure (Just et al., 2010; Just et al., 2014; Mason & Just, 2016).

To evaluate the robustness of the results to the number of voxels used, factor scores from each of the 5 second-order factors were correlated across the different voxel set sizes used in the factor analysis (i.e., 205 voxels, versus the original number of 410, and 615 voxels). The mean correlation between the factor scores from the 410 voxel set (original parameterization) and 615 voxels factor analyses was 0.94 (with all correlations exceeding 0.9). Thus the outcomes are not sensitive to an increase in the numbers of voxels used in the factor analysis. The correlations between the factor scores from the 410 voxel set size (original parameterization) and the 205 voxel set were

somewhat lower: most of the correlations fell to ~0.85 with one of the correlations (corresponding to an Externality/Internality semantic dimension) falling to 0.64. Although the same 5 factors are present when using only 205 voxels, the factor scores are not as similar to the 410 set size. The set of 410 stable voxels was thus used for the factor analysis.

*Predictive Modeling Procedure*

The goal of the predictive modeling procedure was to assess whether the activation pattern of a concept that was left out of the modeling could be predicted, based on the mapping between the behavioral ratings and the activation patterns of all of the other concepts. Accurate predictions would provide converging evidence for the factor interpretations (on which the ratings were based). That is, the correlation between the behavioral ratings of the concepts along the dimensions as we interpreted them and the concepts' factor scores are a test of the interpretation of the factors from the factor analysis. To obtain converging evidence for the factor interpretations, an independent group of participants was asked to rate each stimulus concept on a scale from 1–7 with respect to its salience to the dimensions as they have been interpreted here (e.g., the degree to which a concept, such as faith, was verbally versus perceptually

based). These ratings were then used in a multiple regression model to predict the activation patterns of concepts for which the model had no activation data (Mitchell et al., 2008). Activation predictions for each concept were made within each participant, by developing a separate regression model for each participant to separately predict each concept, basing the model and the weights it derives on the data from the 27 concepts other than the 28th target concept. The resulting model weights were then applied to the dimension ratings and character length of the target concept (Just et al., 2010). These models made predictions of activation values in factor locations obtained from factor analyses that were based on all but the participant in question. The mean prediction accuracies for the 28 concepts were then averaged across participants. A prediction's accuracy was assessed by computing the Euclidean distance between the activation pattern predicted by the model and the observed activation data, relative to the distance to the representations of the other 27 concepts. The normalized rank of the distance between the predicted and test images (among the 28 distances) was the measure of prediction accuracy. Significance was computed using a permutation test. The results of the predicted images with correct labels were compared against the distribution of rank accuracies of predicted images with random labels for 100,000 random permutations.

## Results

### *Systematicity and Commonality of Abstract Concepts*

#### Within-Participant Classification

The mean normalized rank accuracy of the classification of the 28 concepts, first computed for each participant and then averaged over participants, was 0.82, $p < 0.001$ (where chance is 0.5). The mean classification accuracy for each of the 9 participants individually was also reliably above chance (range = 0.76 to 0.94, $p < 0.001$). These results indicate that these abstract concepts have distinctive neural signatures that can be characterized by the multivoxel activation pattern captured by the classifier. Although previous studies have shown that abstract domains such as law and music can be decoded from neural signatures (Anderson, et al., 2014), this finding reveals that individual abstract concepts can be decoded from their neural signatures.

To address the possibility that some of the decoding accuracies could be due to low-level representations of the concept presentation rather than just the concepts, the analyses were repeated excluding left fusiform gyrus (which includes visual word form areas) and bilateral Heschl's gyrus (to account for low-level auditory information) in addition to the previously excluded occipital lobe. The minimal difference between the inclusion and exclusion of these regions (a minor decline from 0.82 to 0.80 in rank

accuracy) suggests that the lower-level word representations have little influence on the overall results.

Representational Similarity between Activation Patterns for Individual Concepts

To explore the similarities among the neural representations of the 28 individual abstract concepts, representational distance matrices (RDMs) were generated using the activation patterns for the 120 most stable voxels for each participant separately. The resulting concept-by-concept RDMs of activation patterns were then averaged across participants. The resulting mean RDM contained 2 clusters of similar concepts. One cluster was related to **mathematics** and **scientific** concepts (top left box of Figure 2.1), including concepts such as subtraction and acceleration. A second cluster indicates similarity of activation patterns among the remaining 5 categories relating to **social**, **emotions**, **law**, **metaphysics**, and **religiosity** (bottom right box of Figure 2.1).

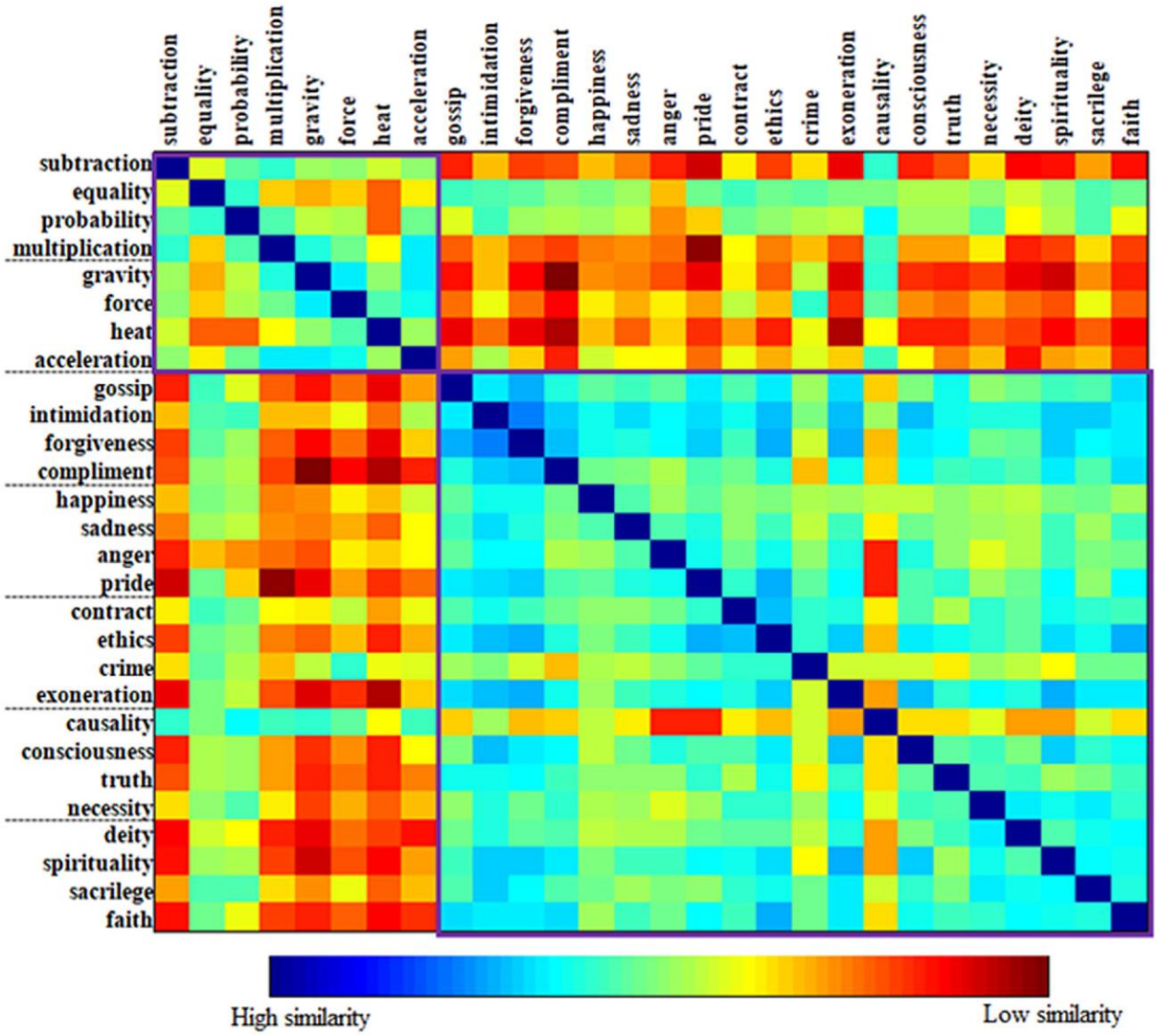


**Figure 2.1.** Representational similarity between neural activation (blue colors indicate higher similarity) for individual concepts. Dotted lines indicate category separation.

## Commonality of Neural Representations across Participants

In addition to establishing that the neural representations of abstract concepts were systematic and decodable within each participant, a between-participant classification was performed to determine whether these abstract concept

representations were similar across participants. When the classifier was trained on the data of all but one participant, the mean rank accuracy for the test data from the left-out participant was 0.74, $p < 0.01$, indicating that the neural signatures had a substantial amount of commonality across participants. All 28 individual concepts were reliably classifiable between participants with a range of 0.58 to 0.94 ($p < 0.01$). Thus these highly abstract concepts are neurally represented as activation patterns that are highly common across participants.

*Factor Analysis for Uncovering Underlying Neurosemantic Dimensions*

A two-level factor analysis (first applied to individual participants, then to the pooled data) was used to uncover the dimensions underlying the activation evoked by the abstract concepts. Four of the five resulting second-level (common) factors that accounted for the most variance were interpretable, using two criteria: 1) the ordering of the 28 concepts by their factor scores for a given factor, particularly the concepts near the two extremes of the ordering; 2) the locations of voxels with high loadings on the factor. These 4 factors were interpreted as corresponding to Verbal Representation, Externality/Internality (to oneself), Social Content, and Word Length. These 4 factors accounted for a total of 33.2% of variance in the group-level factor analyses: Verbal

Representation accounted for 10%; Externality/Internality accounted for 7.9%; Social Content accounted for 6.9%; and Word Length accounted for 8.4%.

Verbal Representation Factor

This dimension is interpreted as the degree to which a concept is represented in verbal as opposed to perceptual terms (Barsalou, 2003). The interpretation of this factor and the others is tested below. This dimension was present for every participant and accounted for the most or second most variance in first-order factor analyses. Concepts with large positive factor scores for this factor included *compliment*, *faith*, and *ethics* while concepts with large negative scores for this factor included *gravity*, *force*, and *acceleration* as shown in Table 2.1.

The main cortical regions containing voxels with high loadings on this factor consisted of LIFG, left anterior supramarginal gyrus (LSMG), and left lateral occipital complex (LLOC; highlighted in red in Figure 2.2). These regions are consistent with a previous meta-analysis examining contrasts between concrete and abstract concepts (Wang et al., 2010). In Wang et al. (2010), GLM contrasts revealed that the areas around LSMG and LLOC activated more for concrete concepts and less for abstract concepts while LIFG activated more for abstract and less for concrete concepts. Even

though the 28 concepts in the present study were all designed to be abstract, the distribution of factor scores along this dimension indicates that some of these concepts, such as *force* and *acceleration* have more perceptual content than others (such as *faith* and *ethics*). The Neurosynth meta-analytic database provides converging evidence for the interpretation of the functional role of LIFG (verbal processing), LSMG (somatosensation), and LLOC (object processing) (http://neurosynth.org; Yarkoni et al., 2011).

**Table 2.1.** Six concepts with the highest and lowest factor scores for each interpretable factor.

| Verbal representation | Externality/internality | Social content | Word length |
|---|---|---|---|
| Compliment (1.78) | Causality (2.41) | Pride (2.11) | Acceleration (1.65) |
| Faith (1.39) | Sacrilege (1.83) | Gossip (1.99) | Exoneration (1.53) |
| Ethics (1.25) | Probability (1.16) | Equality (1.23) | Spirituality (1.52) |
| Truth (1.21) | Deity (1.01) | Forgiveness (1.23) | Multiplication (1.51) |
| Spirituality (1.01) | Gravity (0.84) | Intimidation (1.05) | Causality (1.02) |
| Necessity (0.89) | Equality (0.79) | Gravity (0.8) | Sacrilege (0.98) |
| Subtraction (−0.69) | Pride (−0.94) | Compliment (−1.16) | Pride (−1.04) |
| Causality (−0.87) | Anger (−1.19) | Deity (−1.28) | Faith (−1.14) |
| Heat (−1.74) | Consciousness (−1.38) | Spirituality (−1.5) | Happiness (−1.16) |
| Acceleration (−1.98) | Acceleration (−1.51) | Multiplication (−1.5) | Anger (−1.2) |
| Force (−2.11) | Sadness (−1.72) | Necessity (−1.52) | Crime (−1.49) |
| Gravity (−2.12) | Spirituality (−1.99) | Heat (−1.77) | Heat (−2.02) |

*Note.* Factor scores shown in parentheses.

## Externality/Internality Factor

The second interpretable factor corresponds to the degree to which a concept is experienced as an external versus internal state or event. An event that is external is one

that requires the representation of the world outside oneself and the relative noninvolvement of one's own state. An internal event is one that involves the representation of the self. The main cortical region containing voxels loading on this factor was right supramarginal gyrus (RSMG; see Figure 2.2). This region has been shown to be related to *emotional egocentricity*, that is, "the tendency to project one's own mental state onto others" (Silani et al., 2013). At one extreme of the dimension lie concepts that are external to the self (e.g., *causality*, *sacrilege*, and *deity*). At the other extreme lie concepts corresponding to events that are internal to the participant, such as *spirituality* and *sadness* (Table 2.1). Neurosynth failed to suggest any consistent functional role for the Externality dimension's associated voxel cluster locations. The current interpretation is largely based on the ordering of the concepts by their factor scores on this dimension.

Social Content Factor

A third factor was interpreted to correspond to social content, as it pertains to personal experience. The concepts at one extreme of the dimension included *pride*, *gossip*, and *equality* while the concepts at the other extreme included *heat*, *necessity*, and *multiplication* (Table 2.1). The main cortical region containing voxels with high

loadings for this factor was the left posterior cingulate cortex (LPCC; Figure 2.2), which is associated with the contextualization of oneself in space and emotions (Maddock et al., 2003; Bird et al., 2015; Guterstamet al., 2015). Neurosynth suggests the LPCC is involved in the processing of episodic and autobiographical memories (http://neurosynth.org; Yarkoni et al., 2011). In the context of this study, the LPCC may be involved in the retrieval of memories of previous social interactions.

Word Length Factor

This fourth factor characterizes concepts based on their word length. The concepts that lie on the two extremes of this factor clearly represent the longest and shortest words in the set of concepts. Concepts at one extreme for this factor included *acceleration*, *exoneration*, and *spirituality* while concepts at the other extreme included *heat*, *crime*, and *anger* (*happiness* lying on the "short-word" extreme was an exception). The only cortical region that loaded on this factor was the left occipital pole (Figure 2.2). This finding regarding word-length provides a face validity check for the factor analysis methods and interpretations.

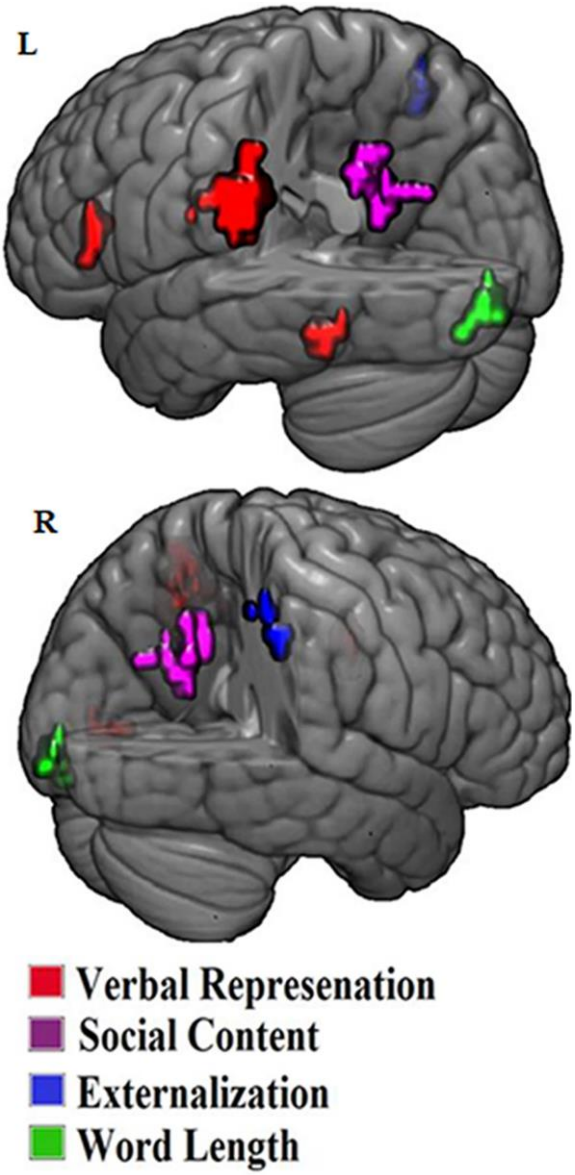


**Figure 2.2.** Locations of the voxel clusters with the highest factor loadings for each of the 4 interpretable factors. Voxels were thresholded to have a minimum cluster size of 15 and mean correlations above 0.2 (in either positive or negative direction) between their activation values and their factor loadings. Cluster centroid XYZ coordinates for: Verbal representation: LIFG (-53.8 22.2 13.4), LSMG (−58.0 −34.1 35.1), and LLOC (−54.3 −62.0 −9.1); Social Content: LPCC (−5.8 −54.0 29.7); Externalization: RSMG (42.9 −41.6 47.4); and Word Length: left occipital pole (−13.0 −96.8 −6.6).

*Testing the Factor Interpretations Using Behavioral Ratings and a Predictive Model*

Ten participants who were not in the fMRI study rated the salience of the 3 semantic factor interpretations to each of the 28 concepts. For example, they rated on a 1–7 scale how verbal (as opposed to perceptually instantiable) items like *gravity* and *ethics* were. The correlation between the mean ratings and the factor scores were 0.63 for Verbal Representation, 0.59 for Externality, and 0.55 for Social Content (all significant at $p < 0.01$), as shown in Figure 2.3). To determine agreement among raters, intraclass correlation was computed for the 3 rated dimensions across participants; ICC was 0.88 for verbal representation, 0.93 for Externality, and 0.97 for Social Content (all significant at $p < 0.01$).

A generative model using the independent ratings (and word length (i.e., number of characters in each word)) was developed to predict the activation of "new" concepts (i.e., concepts left out of the modeling) in the locations corresponding to the factor-associated clusters, based on their association with each of the factors. The mean behavioral ratings served as model weights in a regression model, where the independent variables were the four factors (3 semantic factors and Word Length). To eliminate contamination between the training data that determined the locations and the data, on which the model was tested, the 2-level factor analysis was computed on only 8 participants and the model was tested on the remaining participant. In all 9

iterations of the modeling, the 4 interpretable second-order factors were identified by correlating the factor scores from the 5 second-level factors from the 9-participant factor analysis with each of the 5 second-level factors from the 8-participant factor model. In all iterations, the 4 factors were present with correlations of 0.9 or greater. In each of the 9 iterations of the predictive model, each factor was associated with a set of voxel clusters, and each cluster was characterized by an enclosing cuboid. The 6 most stable voxels were selected from each cuboid of each factor. The mean number of cuboids identified across all 9 iterations are as follows: Verbal Representation contributed a mean of 13.22 (SD= 1.48) cuboids; Externality/Internality contributed a mean of 7.89 (SD= 2.71) cuboids; Social Content contributed a mean of 5.89 (SD= 1.17) cuboids; and Word Length contributed a mean of 2.56 (SD= 1.33) cuboids.

Model predictions were made by leaving out one of the 28 concepts, predicting the activation for that concept using the behavioral ratings (and word length), and computing the Euclidean distance between the predicted activation pattern generated by the model and the observed (test) mean activation data. The normalized rank of the distance between the predicted and test images (among all inter-item distances) was used as a measure of prediction accuracy. This leave-one-out procedure was repeated for all 28 concepts. The mean normalized rank accuracy of the predictions across concepts was 0.78 (SD= 0.09; where chance = 0.5). Mean rank accuracies for all

participants were significantly above chance ($p < 0.001$) as determined using a 100,000-iteration permutation test. Although the factor analysis and its interpretation are exploratory, the correlations between the factor scores and the behavioral ratings, as well as the predictive modeling, provide a clear empirical test of the factor interpretations.

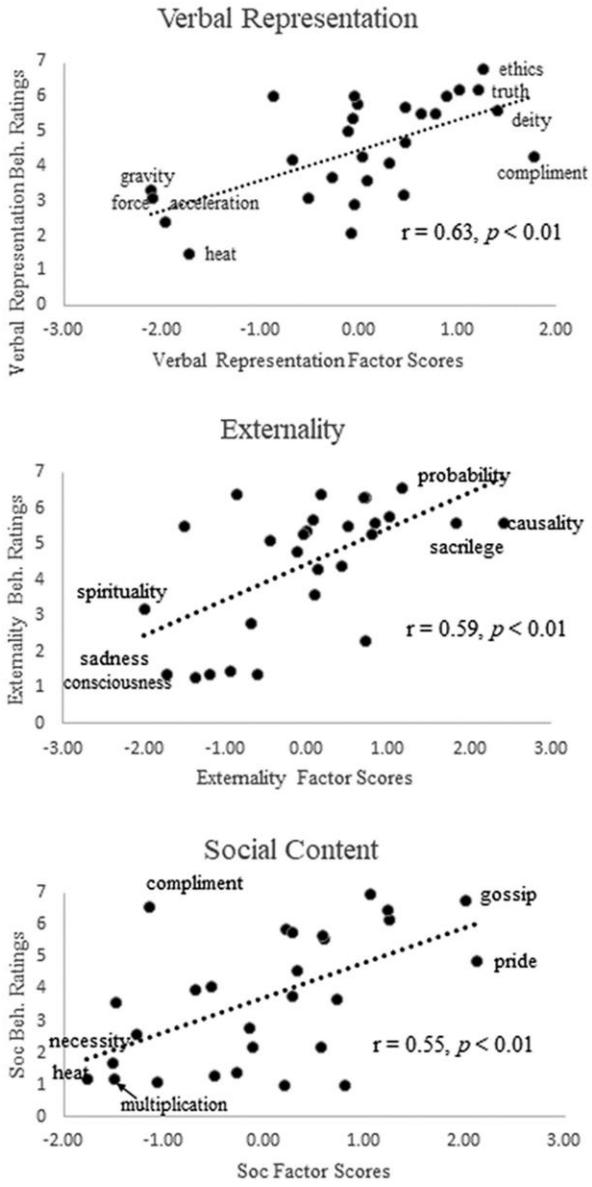


**Figure 2.3.** Scatter plot of factor scores for each of the 3 semantic dimensions versus the mean behavioral ratings of the 28 concepts, and their correlations. Although these correlations are significant, they are based on 28 items and therefore the effect sizes should be interpreted with caution.

*Further Exploration of the Verbal Representation Dimension*

Previous studies have suggested that the relationship between abstractness and activation levels differs for the three regions in the Verbal Representation factor (i.e.,

LIFG, LSMG, and LLOC; Wang et al., 2013). Correlations between the second-level factor scores from this dimension and the MPSC activation levels were computed for each voxel in these 3 subregions for each participant separately. The correlations values were then averaged over the participants within each voxel. LIFG activated more for concepts that are more verbally represented ($r = 0.38$, $p < 0.05$) whereas LSMG and LLOC activated more for concepts that are more perceptually represented ($r = 0.46$, $p < 0.05$), as shown in Figure 2.4. These results suggest that the abstractness of a concept corresponds to the degree to which it is represented in verbal terms, which can be thought of as a point along a verbal-perceptual continuum. To further investigate whether one of the variables underlying the Verbal Representation factor is concreteness, the 28 concepts' factor scores for this dimension were compared with their concreteness ratings from Brysbaert et al. (2014), resulting in a substantial correlation ($r = -0.47$, $p < 0.05$).

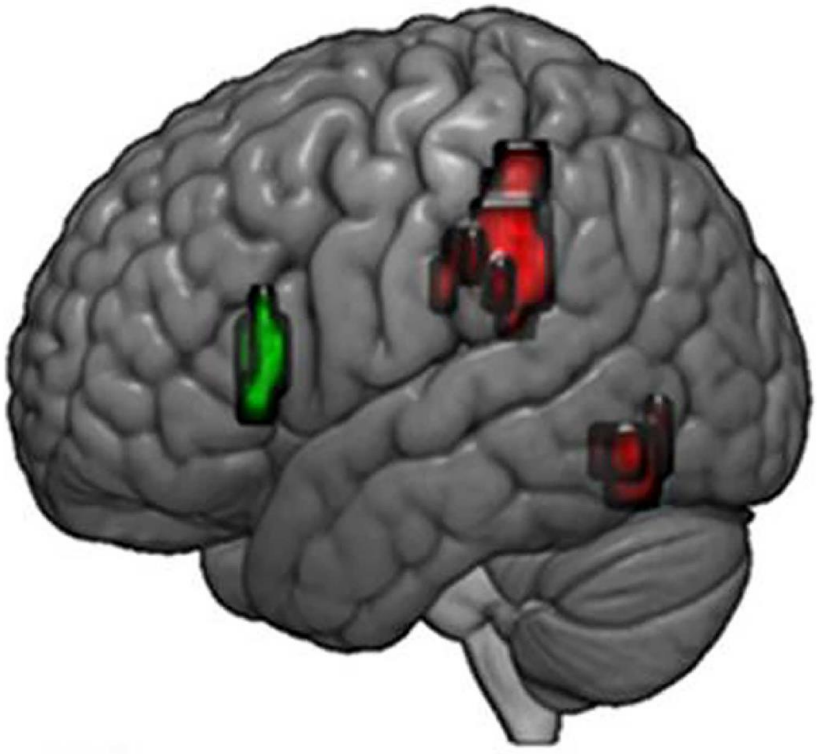

**Figure 2.4.** Correlation between the Verbal Representation factor scores and MPSC activation. As concepts become more verbally represented they recruit more LIFG and show less activation in regions associated with the visual representation of concepts (LSMG and LLOC). Positive correlations shown in green; negative correlations in red.

## Discussion

The human ability to think about abstract entities plays a central role in scientific and intellectual progress. The ability to deeply understand the nature of the world around us (including the sociopolitical world) depends on the repeated application of this ability over millennia. Despite the intuitive consensus of which concepts are abstract, it was not known what neurally characterizes an abstract concept, beyond its

preferential recruitment of left frontal language-based areas (Binder et al., 2005; Wang et al., 2010).

The primary results of this study can be summarized as follows: first, there is enough consistent and common information in the neural signatures of abstract concepts to reliably identify a set of 28 such concepts within and across participants. Second, the neural representations of these concepts are underpinned primarily by 3 semantically interpretable dimensions (*Verbal Representation*, *Externality/Internality*, and *Social Content*). Third, the abstractness of a concept is defined not only by the absence of concreteness but also in terms of its verbal characterization. This study provides new insight into the neural systems and underlying implicit semantic structures that are used to represent abstract concepts.

*Systematicity and Commonality of Abstract Concepts*

Given the absence of common perceptual content related to abstract concepts, there was reason to anticipate substantial individual differences in the representations of such concepts. Nevertheless, the between-participant classification was reliably accurate, indicating considerable nonperceptual commonality in the meaning representations. The variation among the concepts in their across-participant

classification accuracy provides hints at what makes an abstract concept representation less or more common. Concepts such as *anger* and *multiplication* were less well predicted than others across participants (although still reliably so), and these concepts tended to be highly instantiable. By contrast, concepts such as *necessity*, which are highly verbally represented, were extremely well predicted across participants. Thus a post hoc hypothesis is that across-participant commonality is greater for more verbally-based concepts and somewhat lower for more instantiable concepts, which may be instantiated differently across participants.

### *Semantic Primitives Associated with the Neural Representation of Abstract Concepts*

The three semantic dimensions underlying the representation of abstract concepts are Verbal Representation, Externality, and Social Content. That is, we propose that abstract concepts are represented based on: their meaning across a wider variety of contexts than concrete concepts (Crutch & Warrington, 2005, 2010; Hoffman, 2016); their reliance on using the self as a reference point; and their use of social contexts as a reference point. It is useful to highlight that these representations of abstract concepts were based on neural activation patterns. It is possible to assess semantic representations of abstract concepts based on different types of data, such as

cooccurrence properties in large corpora or behaviorally measured semantic features (Wang et al., 2017). The dimensions identified in this study provide a neurally-driven foundation for understanding the semantic underpinning of abstract concepts.

The factor analysis procedure identifies regions reflecting the organization of the 28 concepts along various dimensions. However, none of the factor locations included the anterior temporal lobe (ATL), which has been shown to activate to both concrete and abstract words (Jefferies et al., 2009; Hoffman, 2016) and has also been shown to be involved in the integration of low-level perceptual features of visual objects (Coutanche & Thompson-Schill, 2015). A GLM contrast of the 28 abstract concepts vs. fixation revealed activation in the superior portion of the ATL, indicating that ATL may serve a similar function for all 28 abstract concepts.

One of the strengths of the approach that was used here is the quantitative assessment of the fit of the interpretation of each dimension to the activation data. Although the interpretations fit the data well, as with any theoretical proposal, alternative interpretations can be generated and quantitatively assessed.

*Degree of Abstractness as a Point on a Gradient between Language and Percepts*

The Verbal Representation factor organizes conceptual representations based on the dissociation of activity in neural structures associated with verbal processing (LIFG) and spatial/object processing (Figure 2.4; Grill-Spector et al., 2001). LIFG has been reliably shown to be involved in verbal processing (Yarkoni et al., 2011; Hoffman, 2016). It is incomplete to say that the abstract concepts evoke less activation in regions associated with perceptual processing; rather, abstract concepts both evoke less activation in regions associated with perceptual processing and evoke more activation in regions strongly associated with verbal processing. This dissociation in neural patterning suggests that the degree of abstractness of a concept is a point on a continuum between language systems and perceptual processing systems. This result provides a neural realization for the intuitive idea that abstractness is not a binary construct but rather a gradient-like translation of a concept into amore verbal encoding.

This point raises an interesting theoretical question regarding the role of neural language systems, particularly LIFG, in the verbal representation of abstract concepts. LIFG has been implicated in the integration of semantic relationships among different contexts. Abstract concept representations require an integration of meaning from a greater variety of contexts relative to concrete concepts (Crutch & Warrington, 2005, 2010; Hoffman, 2016; Hayes & Kraemer, 2017). Thus, LIFG may become more activated

for the concept ethics than gravity because ethics requires integration across more semantically variable contexts. The activation in LSMG (and LLOC), by contrast, is related to the instantiability of a concept (Figure 2.4). The critical finding here is that the degree of perceptual involvement varies systematically across abstract concepts.

## Conclusion

The lack of a perceptual grounding makes abstract concepts difficult to characterize in semantic and psychological terms, but a neural framework provides a good beginning to the answer. What neurally defines the abstractness of a concept is its place on a continuum between perceptible experience and a purely verbal entity. This continuum emerges even among a set consisting entirely of abstract concepts. Moreover, the present study suggests that abstract concepts rely on semantic features that are also not necessarily perceptually grounded, such as our ability to construe abstract concepts relative to ourselves, or to use social contexts as a reference.

# Chapter 3. Project 2: Similarities and differences in the neural representations of abstract concepts across English and Mandarin

Chapter 2 provided a published empirical examination of 28 abstract concepts for speakers of the same language. Three semantic dimensions describing the underlying structure of abstract concepts were identified including: **Verbal representation**; **Internality/Externality**; and **Social**. These dimensions provide a framework for understanding how abstract concepts are represented in the brain. However, the question of the generalizability of these dimensions remains. Chapter 3 describes a published empirical study examining the neural representation of the same set of 28 concepts across speakers of different languages, specifically, American native English speakers and Chinese native Mandarin speakers. Using similar MVPA methodologies, the underlying structure of abstract concepts will be assessed across both languages.

## Introduction

Although the neural representations of concepts are generally similar across speakers of the same language, the extent of this similarity across languages has yet to be measured. When the concept corresponds to a concrete entity, such as an apple, the common basis in large part consists of the perceptual and physical properties of the

referent (Just et al., 2010). Recent studies using multivariate pattern analyses (MVPA) and machine learning techniques have reported cross-language decoding of fMRI signatures, namely, across English and Portuguese nouns, (Buchweitz et l., 2012), across English, Portuguese, and Mandarin sentences (Yang et al., 2017a, 2017b) as well as English, Mandarin, and Farsi stories (Dehghani et al., 2017).

The shared representational basis of abstract concepts such as ethics and causality are more difficult to identify. Given that abstract concepts do not often reflect a shared experience of the physical world, require schooling to acquire (Mason & Just, 2016), and are built on existing conceptual knowledge, there is reason to question the degree of commonality across languages in the meaning representations underlying abstract concept knowledge. Some theories have suggested that the psychological representations of abstract concepts, such as time, are dependent on cultural and language differences (Fuhrman et al., 2011; Lai & Boroditsky, 2013) while other theories suggest that there are culturally-invariant neural activation patterns for concepts across brain regions (Han & Northoff, 2008). Although abstract concepts have been shown to be represented similarly across speakers within a given language (Vargas & Just, 2020), it has yet to be measured whether or not this common representation extends across languages.

Among English speakers, the neural activation patterns for abstract concepts have been shown to be underpinned by a set of three neurosemantic dimensions, namely the degree to which a concept is verbally represented; whether a concept uses the self as an internal reference; and whether the concept contains social content. Furthermore, in English, the neural representation of abstract concepts has been shown to involve regions associated with motor and visuospatial functioning (Dreyer & Pulvermüller, 2018; Harpaintner et al., 2020). Other research has supported the emphasis of verbal and linguistic-based processing of abstract concepts in Mandarin speakers (Wang et al., 2018). The current study compared the neural representations of the same abstract concepts in English and Mandarin to illuminate commonalities and possible differences between languages in the representation of abstract concepts.

This study had two main aims: First, to test whether a shared set of semantic dimensions underlie the neural activation patterns of abstract concepts across English and Mandarin speakers and to determine how well the observed organization of abstract concepts along these dimensions corresponds to behavioral judgments of concept meaning. Second, to identify differences in the representation of individual concepts despite a common underlying structure. Taken together, this study aims to determine whether there is a common neural basis for representing abstract concept information

across languages while providing a framework for identifying language-specific differences in the meaning of individual abstract concepts.

## Methods

### *Participants*

Ten right-handed native Mandarin speaking adults (age range from 18 to 26, M = 20.2; six females) and 10 right-handed native English-speaking adults (sample previously reported in Vargas & Just, 2020) age range from 20 to 38, M = 25.9; seven females;) from the Carnegie Mellon community participated in a 45-min fMRI scanning session. To mitigate cross-cultural familiarity, the group of native Mandarin speakers included only those who had spent less than 1 year living outside of the Peoples Republic of China. Informed consent was obtained from all participants in accordance with the Carnegie Mellon Institutional Review Board. Data from two Mandarin speakers and one English speaker were excluded due to the participant falling asleep during the scan. An additional Mandarin speaking participant's data was excluded due to their misunderstanding of instructions, resulting in data analysis of seven Mandarin speakers and nine English speakers.

*Experimental Paradigm*

For both language groups, the stimuli were 28 words referring to abstract concepts distributed among seven categories. Although the category labels were never mentioned nor presented to participants, they are listed here in parentheses for expository purposes, preceding the actual stimuli: (**social**): *gossip*, *intimidation*, *forgiveness*, and *compliment*; (**emotion**): *happiness*, *sadness*, *anger*, and *pride*; (**law**): *contract*, *ethics*, *crime*, and *exoneration*; (**metaphysics**): *causality*, *consciousness*, *truth*, and *necessity*; (**religiosity**): *deity*, *spirituality*, *sacrilege*, and *faith*; (**mathematics**): *subtraction*, *equality*, *probability*, and *multiplication*; (**scientific**): *gravity*, *force*, *heat*, and *acceleration*. The set of concepts was translated from English to Mandarin by two independent native Mandarin speakers and then back-translated to English by a separate independent Mandarin speaker. The translations were then verified by a fourth independent Mandarin–English bilingual to ensure the meaning best matches the original English concept (see Table 3.1 for Mandarin translations).

**Table 3.1.** Table of all 28 abstract concepts stimuli presented to English and Mandarin speaking participants

| Math | Scientific | Social | Emotion | Law | Metaphysical | Religiosity |
|---|---|---|---|---|---|---|
| Subtraction (减法) | Gravity (引力) | Gossip (绯闻) | Happiness (幸福) | Contract (合同) | Causality (因果关系) | Deity (神明) |
| Equality (相等) | Force (力) | Intimidation (恐吓) | Sadness (悲伤) | Ethics (道德) | Consciousness (意识) | Spirituality (灵性) |
| Probability (概率) | Heat (热能) | Forgiveness (谅解) | Anger (愤怒) | Crime (罪行) | Truth (真理) | Sacrilege (亵渎) |
| Multiplication (乘法) | Acceleration (加速度) | Compliment (赞美) | Pride (自豪) | Exoneration (免罪) | Necessity (必要性) | Faith (信仰) |

*Note*: Stimuli were presented in the participant's native languages.

Concept abstractness ratings were compared across the languages. English abstractness ratings were obtained from the Brysbaert et al. (2014) database while Mandarin ratings were obtained from MELD-SCH (Xu & Li, 2020). Because the concepts in the MELD-SCH were limited to words with two characters, abstractness comparisons were restricted to the 18 concepts present in both databases (18 of 28 concepts), $r = 0.64$, $p < 0.01$. Word frequencies were compared across languages using English (Brysbaert et al., 2014) and Mandarin (Cai & Brysbaert, 2010) word frequency databases. The correlation comparing the word frequencies of the concepts across languages was $r = 0.3$, $p = 0.12$, indicating some minor differences in word frequency across languages.

Prior to the scanning session, participants were presented with a list of the 28 concepts and asked to write down three prominent properties of the concept's meaning. Possible properties included synonyms, definitions, or experiences associated with the

concept intended to guide participants to mentally evoke a consistent representation for each concept. Participants were instructed to write properties that came to mind quickly and naturally.

There was a total of six presentation blocks of the same 28 stimulus concepts (using different random permutation orders in the different presentations) in the scanning session, distributed between three runs (two blocks per run) to allow participants a brief rest between runs. A 17 s "*X*" was presented at the beginning of each block (two per run) to use as a baseline measure of neural activity. The set of 28 stimuli was presented six times to provide multiple datasets for training and testing the machine learning classifier in its cross-validation protocol. Prior to the scan, participants briefly practiced the experimental paradigm in a mock MRI scanner while receiving head-motion feedback to minimize movement.

On each trial, participants were visually presented with the stimulus word concept in their native language for 3 s and were asked to think about the properties associated with that concept. Following this 3 s period, participants were instructed to clear their mind over the course of 7 s while watching a blue ellipse shrink to nonexistence, to allow the hemodynamic response to approach baseline before the next concept appeared. The shrinking ellipse provided a visual fixation target and conveyed the progress through the 7 s interstimulus interval.

*fMRI parameterization and image processing*

Functional images were acquired on a Siemens Verio 3.0T scanner and a 32-channel phased-array head coil (Siemens Medical Solutions, Erlangen, Germany) at the Scientific Imaging and Brain Research facility (SIBR) at Carnegie Mellon. Scans were acquired using a gradient-echo echo-planar imagining pulse sequence (TR = 1,000 ms, TE = 25 ms, and a 60˚ flip angle); each volume contained 20 5-mm thick AC-PC aligned slices (1-mm gap between slices). The acquisition matrix was 64 x 64 with 3.125 x 3.125 x 5-mm voxels. SPM8 (http://www.fil.ion.ucl.ac.uk/spm/) was used to correct for head motion and normalize to the Montreal Neurological Institute template. The percent signal change (PSC) relative to the fixation condition was computed at each gray matter voxel for each stimulus presentation (the PSC data was converted to z-scores).

The main measure of activation evoked by a concept consisted of the voxel activation levels acquired around the peak of the hemodynamic BOLD response, namely the mean of four brain images acquired once per second (i.e., a TR of 1,000) within a 4 s window, offset 5 s from the stimulus onset (i.e., images 5–8). Mean PSCs were normalized across voxels for each trial (MPSC). Previous studies have reported that the mean activation across these four images (as opposed to a GLM measure) yields a high

classification accuracy obtained by a classifier that relates the activation pattern to the concept (Bauer & Just, 2017; Just et al., 2010; Mason & Just, 2016).

*Voxel Stability*

The analysis focused on the most stable voxels, those whose activation levels were similarly modulated by the set of 28 abstract concepts each time the set was presented. This property selects voxels whose activation levels constitute neural signatures of a set of concepts (Bauer & Just, 2017; Just et al., 2010, 2017; Kassam et al., 2013; Mason & Just, 2016; Mason & Just, 2020; Yang et al., 2017a, 2017b). Thus, a voxel with high stability is one that has a stable tuning curve over the set of stimuli. A voxel's stability was computed as the mean pairwise correlation of its 28 MPSC activation levels (for the 28 abstract concepts) across all pairwise combinations of the presentation blocks in the training data. Stable voxels were used as features in classification and factor analyses. The stable voxels selected in the training data for classification are then used in the test set. The 120 most stable voxels in the whole brain were used as features for classification. This approximate number of voxels has been shown to reliably capture meaningful information in the neural representation of individual concepts (Just et al., 2010; Mason & Just, 2016). To demonstrate that the results and conclusions are not

particularly sensitive to variations in the number of features, the classification analysis was repeated varying the number of stable voxels used from 20 to 10,000 (in 20 voxel increments); the peak classification accuracy occurred between 120 and 180 stable voxels. The mean classification accuracy gradually decreased with the inclusion of additional stable voxels beyond 180. To be consistent with previous studies, 120 stable voxels were used as features.

*Within-participant Classification*

The data were analyzed using various classification approaches, each informing a different aspect of the underlying concept representations. Within participant concept classification captures participant specific reliability as well as idiosyncrasies in concept representations. High accuracies in the within participant classification analyses suggest individual participants were able to think about a specific concept consistently and distinctly, making them identifiable by the classifier. A Gaussian Naïve Bayes (GNB) classifier was trained to decode the 28 concepts, based on its training on an independent subset of the activation data from four of the six presentations and it was tested on the mean of the two left-out presentations. This cross-validation procedure was followed in 15 (six choose two) folds. The features used by the classifier consisted of the activation

levels of the 120 most stable voxels in the training set from anywhere in the whole brain. The classifier's mean normalized rank accuracy was used to assess decoding accuracy (i.e., the mean over folds of the normalized rank of the correct response in a probability-ranked list of all 28 alternatives, where chance level is 0.5). Chance performance was determined using a 10,000-iteration permutation test on each participant separately for each concept-level prediction.

*Between participant, within language classification*

Between participant within-language classification quantifies the commonalities of the neural representations across participants of the same language. For each language group separately, a GNB classifier was trained on the neural signatures of the concepts from all but one participant and tested on the left-out participant's data. The mean rank accuracy was computed across the resulting nine folds for the English group and seven folds for the Mandarin group. Chance performance was determined using a 10,000-iteration permutation test. The voxels used in the classification across participants were those with the highest stability across participants from that participant's language group. To compute the cross-participant stability of voxels, the MPSC data was first averaged across all presentations for each participant, and then the

mean pairwise correlation of a voxel's 28 MPSC activation levels (for the 28 abstract concepts) was computed between all pairs of the remaining participants in the training data. The 120 most stable voxels (i.e., those with the highest mean pairwise correlation) from the whole brain across the training participants (eight for the English group, six for the Mandarin group) were selected as features for the classifier. The methods for the cross-language classification, which was based on the factor locations, are described below after the factor analyses.

*Factor Analysis*

To uncover the semantic dimensions underlying the representations of the 28 abstract concepts, a two-level factor analysis was computed based on the combined data from the participants of both languages; a factor analysis was first applied to the data of individual participants and then the second factor analysis used the factor scores from the first level as input (using a procedure described in detail in Just et al., 2014). The factor analysis of the English-specific activation data was previously reported and used similar methods (Vargas & Just, 2020). The factor analysis of the Mandarin-specific data followed the same procedure with the exception that 6 second-level factors were

extracted instead of five. The factor analysis was implemented using a principal factor analytic algorithm in MATLAB (R2011a; version 7.12; The MathWorks, Natick, MA).

The inclusion of brain regions in the combined-language second level factor analysis was based on broad AAL (Automated Anatomical Labeling) regions containing voxels that met three criteria: the voxels had to: (1) be stable in the cross-participant stability map; (2) have factor loadings above a threshold of ≥ 0.4; and (3) form clusters of at least 15 contiguous voxels. Spheres were then generated using the centroids of these clusters. The data from all 16 participants (seven Mandarin and nine English) were analyzed to identify interpretable factors. As described in Vargas and Just (2020), an initial map of the union of 800 stable voxels from each language was generated. This map was then parcellated using AAL (Tzourio-Mazoyer et al., 2002). The parcellated map was then used to identify AAL-defined regions with large numbers of stable voxels relative to the total number of voxels in the AAL region. Then, the input to the first-level factor analysis (performed within each participant) consisted of the mean activation levels of the most stable voxels in each of the contributing AAL regions. The total number of voxels used in this factor analysis was 410, similar to the number used in previous studies (Kassam et al., 2013; Vargas & Just, 2020). The 410 voxels were selected with the number per AAL-defined ROI based on the numerosity of the ROI's stable voxels in the initial map: 40 voxels from left inferior frontal gyrus (LIFG); 30

voxels from left posterior cingulate cortex; 60 voxels from frontal cortex bilaterally; 60 voxels from occipital cortex bilaterally; 60 voxels from temporal cortex bilaterally; and 160 voxels from parietal cortex bilaterally. Because the results have been shown to be insensitive to minor variations in the data analysis parameters, the same parameter values were used in this study as in Vargas and Just (2020). To assess the dependency of the analyses on the choice of particular parameter values, the combined-language second-level factor analyses were computed with systematic variation of several parameters, namely the number of input voxels, number of first level factors, and number of second level factors. The effects of these variations were evaluated by correlating the factor-scores from the second-level dimensions across the variations and comparing the locations of the voxel clusters with high factor loadings across the variations. The effects of these variations were found to be minor, so the parameter values used in the analysis of this study were the same as those used in the previous study of these concepts (Vargas & Just 2020).

This first-level factor analysis was performed on all 16 participants individually, extracting seven factors for each subject, resulting in a total of 112 vectors of factor scores. A voxel was determined to belong to a factor if its factor loading exceeded a threshold 0.4 (a typical value for a factor loading threshold). This same threshold was used in previous studies that characterized brain locations identified through factor

analysis (Just et al., 2010, 2014; Mason & Just, 2016). To eliminate isolated single voxels, the factor-loading voxels were required to form clusters containing a minimum of 15 voxels. Spheres for each factor were generated based on the centroids of clusters and extend to account for minor inter-participant variations in specific voxel locations for that factor.

The goal of the first-level factor analyses was to partition the set of input voxels into subsets that responded similarly across the set of abstract concepts, specifying seven factors. This analysis produced factor scores for the 28 concepts, for each of the seven factors, for each of the 16 participants. The 16 participants' seven sets of factor scores were concatenated and used as input into the second, group-level factor analysis (a total of 112 sets of 28 factor scores) to further reduce the dimensionality to six dimensions and to seek consistency across participants and languages. To evaluate the robustness of the factor results, analyses were computed with varying number of input voxels and factors. Although there were minor variations in the scores of individual concepts, the overall factor interpretation and factor scores for concepts remained generally unchanged.

To confirm there is a common neural basis across languages, a factor analysis was computed on both languages separately and the factor scores were correlated between languages for each identifiable dimension. The correlations for the second-level factor

scores across languages for each identifiable dimension are as follows: Verbal representation: r = 0.55, $p < 0.01$; rule-based: r = 0.45, $p < 0.05$; social content: r = 0.42, $p < 0.05$; externality–internality: r = 0.32, $p < 0.1$; word length: r = 0.20, n.s. Notably, the previously unexplainable factor described in Vargas and Just (2020) was reliably correlated with the newly identified rule-based factor in the Mandarin group. Additionally, the lack of correlation between the low correlation of the word length factor scores across languages reflects language-specific orthographic differences.

Regions for the factor analysis of both languages combined were selected based on their being populated by stable voxels. Whole-brain voxel-wise stability was computed for each participant separately and averaged across participants. This method allows for a spatially stable common set of voxels to be identified. The interpretation of each individual factor was largely based on the distribution of the corresponding factor scores across the 28 concepts (particularly the nature of the items at the two extremes of the factor scores) and based to some degree on previous findings that associated particular processes with the factor locations. Moreover, converging evidence for the factor interpretations was provided by the correlation between the factor scores and independent participant ratings of the items with respect to the factor as interpreted.

*Behavioral rating of semantic dimensions*

To obtain converging evidence for the factor interpretations, an independent group of 20 participants (10 native English speakers and 10 native Mandarin speakers) were asked to rate each stimulus concept on a scale from 1 to 7 with respect to its salience to the dimensions as they were interpreted here (e.g., the degree to which a concept, such as ethics, was verbally vs. perceptually based). These ratings of the concepts along each of its dimensions were then used as independent variables in a multiple regression model to predict the activation pattern of a concept in the factor locations (Just et al., 2010; Vargas & Just, 2020).

The correlations between the behavioral ratings of English and Mandarin participants for the 28 concepts on each dimension were as follows: Verbal representation, $r = 0.67$, $p < 0.001$; externality–internality, $r = 0.93$, $p < 0.001$; rule-based, $r = 0.94$, $p < 0.001$; social content, $r = 0.9$, $p < 0.001$. Given the highly reliable correlation between English and Mandarin behavioral ratings, averaged ratings were used as input to the regression model.

*Predictive modeling*

To evaluate the how well the factor interpretations fit the activation data, a predictive modeling procedure was used to assess whether the activation pattern of an individual concept could be predicted, based on the mapping between behavioral ratings of all the other concepts in the set (i.e., leaving out the to-be predicted item) with respect to the factor interpretations and their activation patterns. Accurate predictions would provide face validity for the factor interpretations. Activation predictions for each concept were made by developing a separate regression model for each participant to predict a left-out concept's activation pattern, based on the model weights from the remaining 27 concepts. The factor locations used were obtained from factor analyses based on all participants except for the one being predicted. The mean prediction accuracies for the 28 concepts were then averaged across participants. A prediction's accuracy was assessed by computing the Euclidean distance between the activation pattern predicted by the model and the observed activation data, relative to the distance to the representations of the other 27 concepts. The normalized rank of the distance between the predicted and test images (among the 28 distances) was the measure of prediction accuracy. Significance was computed using a permutation test. The results of the predicted images with correct labels were compared against the distribution of rank accuracies of predicted images with random labels for 10,000 random permutations.

*Factor-based cross-language classification*

Cross-language factor-based classification quantifies the commonality of representation across languages based on the semantic dimensions underlying the concept representations. To test whether the factors (or dimensions) are sufficient for identifying the neural signatures of individual abstract concepts across languages, a GNB classifier was trained on the neural signatures from all participants from one language and was tested individually on each of the participants from the other language. The data consisted of the mean MPSC values of each concept across repetitions for each participant in the factor locations of the five interpretable factors in the factor analysis including both languages. A classifier was trained on the data of all nine native English speakers and was tested on each of the seven native Mandarin speakers and vice versa. The 28 rank accuracies from each participant in the test language were then averaged. There were minimal differences in accuracies between the two classifiers $t(27) = 0.01$, n.s., so the accuracies of the two classifiers were averaged. Above-chance performance at $p < 0.01$ is 0.56 for concept-level predictions as determined using a 10,000-iteration permutation test.

## Results

*Commonality of abstract concept representations within and across languages*

Within-participant classification in the two languages

The individual 28 abstract concepts were reliably identified from their multi-voxel neural signatures within each language by a classifier. This mean classification accuracy for native English participants, 0.83, was reliably above chance (range = 0.76–0.94, $p < 0.001$; mean cutoff for $p < .001 = 0.60$; SD = 0.003) as was that of the seven native Mandarin participants (mean = 0.77; range = 0.66–0.84). Although the concepts of all participants in both groups were identifiable, a t-test comparing the within-participant classification accuracies of the 28 concepts across languages indicated that the classification accuracy was reliably higher in the English participants, $t(27) = 6.70$, $p < 0.001$.

Although the concepts differ in their overall identifiability between the two language groups, these results indicate that these abstract concepts have distinctive neural signatures in both languages that can be characterized by the multi-voxel activation pattern captured by the classifier.

Commonality of the concept representations across speakers of the same language

A between-participant, within-language classification was performed to determine whether these abstract concept representations were similar across speakers within a language group. For English speakers, when the classifier was trained on the data of all but one participant, the mean rank accuracy of the concept identification in the data from the left-out participant was 0.74, $p$ <.01, indicating that the neural signatures had a substantial amount of commonality across participants (Table 3.2). All 28 individual concepts were reliably classifiable between English-speaking participants, with a range of 0.58–0.94 ($p$ < 0.01=0.55). For Mandarin speakers, when the classifier was trained on the data of all but one participant, the mean rank accuracy of the concept classification in the test data from the left-out participant was 0.73, $p$ < .01 (Table 3.2). All 28 individual concepts were reliably classifiable between Mandarin-speaking participants, with a range of accuracies from 0.57 to 0.92 ($p$ < 0.01 = 0.54) except for contract which was classifiable only at $p$ < .05. Thus, there is a comparable degree of commonality across participants within each language group in their neural representation of the abstract concepts. Below, the underlying dimensions of the concept representations across languages are described, followed by an assessment of the commonality of the neural representations of individual concepts across languages, taking the underlying dimensions into account. In the few cases where between-

participant decoding was more accurate than within-participant decoding within a language, the difference might be attributable to the different way the stable voxels were selected in the two cases. The consensually chosen stable voxels in the between-participant analysis could have reduced idiosyncratic properties in the concept representations.

**Table 3.2.** Commonality of concepts within and across languages as measured using concept-level decoding rank accuracy. Dashed lines separate concept categories.

| | Mandarin Between-Participant | Mandarin Within-Participant | English Between-Participant | English Within-Participant | Cross-Language |
|---|---|---|---|---|---|
| Subtraction | 0.72 | 0.79 | 0.94 | 0.89 | 0.70 |
| Equality | 0.7 | 0.77 | 0.58 | 0.8 | 0.67 |
| Probability | 0.69 | 0.72 | 0.8 | 0.86 | 0.49 |
| Multiplication | 0.91 | 0.81 | 0.73 | 0.92 | 0.82 |
| Gravity | 0.8 | 0.8 | 0.86 | 0.88 | 0.79 |
| Force | 0.84 | 0.84 | 0.84 | 0.87 | 0.83 |
| Heat | 0.63 | 0.7 | 0.78 | 0.8 | 0.74 |
| Acceleration | 0.92 | 0.77 | 0.76 | 0.84 | 0.77 |
| Gossip | 0.63 | 0.79 | 0.66 | 0.79 | 0.68 |
| Intimidation | 0.72 | 0.82 | 0.75 | 0.83 | 0.73 |
| Forgiveness | 0.75 | 0.69 | 0.77 | 0.7 | 0.80 |
| Compliment | 0.81 | 0.72 | 0.67 | 0.81 | 0.58 |
| Happiness | 0.76 | 0.71 | 0.75 | 0.76 | 0.60 |
| Sadness | 0.69 | 0.78 | 0.71 | 0.85 | 0.76 |
| Anger | 0.59 | 0.79 | 0.62 | 0.82 | 0.62 |
| Pride | 0.78 | 0.86 | 0.86 | 0.88 | 0.84 |
| Contract | 0.52 | 0.77 | 0.63 | 0.75 | 0.66 |
| Ethics | 0.71 | 0.78 | 0.72 | 0.83 | 0.56 |
| Crime | 0.66 | 0.72 | 0.76 | 0.8 | 0.69 |
| Exoneration | 0.63 | 0.77 | 0.76 | 0.83 | 0.41 |
| Causality | 0.91 | 0.89 | 0.8 | 0.89 | 0.62 |
| Consciousness | 0.66 | 0.77 | 0.79 | 0.84 | 0.53 |
| Truth | 0.69 | 0.77 | 0.62 | 0.86 | 0.55 |
| Necessity | 0.9 | 0.82 | 0.78 | 0.77 | 0.57 |
| Deity | 0.77 | 0.69 | 0.59 | 0.78 | 0.60 |
| Spirituality | 0.57 | 0.72 | 0.79 | 0.82 | 0.46 |
| Sacrilege | 0.61 | 0.73 | 0.62 | 0.83 | 0.52 |
| Faith | 0.81 | 0.72 | 0.78 | 0.81 | 0.61 |
| Mean | 0.73 | 0.77 | 0.74 | 0.83 | 0.65 |

## Mandarin-specific factor analysis

The Mandarin-specific factor analysis indicated a common neurosemantic basis for the set of 28 abstract concepts across English and Mandarin, revealing five interpretable dimensions, namely: Verbal representation, social content, rule-based, externality/internality, and word length. The concepts located at the extremes of each of

these dimensions and their respective factor scores are shown in Table 3.3. The correlations between the Mandarin behavioral ratings and Mandarin-only factor scores for the 28 concepts on each dimension were as follows: Verbal representation, r = 0.57, $p < 0.01$; externality–internality, r = 0.66, $p < 0.001$; rule-based, r = 0.34, $p = 0.07$; social content, r = 0.27, $p = 0.16$. The brain locations of the voxel clusters with high loadings on the interpretable factors for the Mandarin-specific analysis are shown as spheres in Figure 3.1.

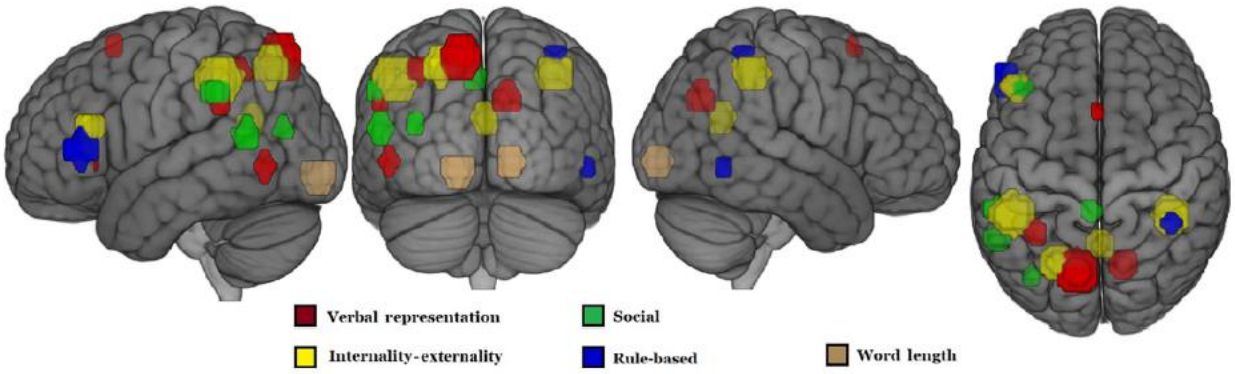


**Figure 3.1.** Locations for five interpretable factor dimensions from Mandarin-specific analysis. These spheres were specified using the centroids of clusters of voxels (containing a minimum of 10 voxels) with high loadings (>0.4) on each of the factors.

**Table 3.3.** Mandarin-only factor analysis output including: six concepts with the highest and lowest factor scores for each mapped dimension, factor locations, and correlations between factor scores and behavioral ratings.

| Verbal representation | Externality/internality | Rule-based | Social content | Word length |
|---|---|---|---|---|
| Faith (2.17) | Subtraction (2.20) | Truth (1.48) | Intimidation (1.73) | Causality (3.23) |
| Spirituality (1.78) | Equality (1.84) | Acceleration (1.46) | Sadness (1.63) | Necessity (2.22) |
| Deity (1.24) | Gravity (1.57) | Gravity (1.03) | Equality (1.52) | Acceleration (1.61) |
| Compliment (1.19) | Force (1.00) | Sadness (1.02) | Contract (1.21) | Happiness (0.89) |
| Probability (1.02) | Causality (0.93) | Causality (0.96) | Ethics (1.06) | Exoneration (0.65) |
| Causality (0.98) | Contract (0.50) | Force (0.84) | Gossip (0.65) | Compliment (0.46) |
| Forgiveness (−1.10) | Spirituality (−0.78) | Consciousness (−1.34) | Crime (−0.78) | Sadness (−0.72) |
| Consciousness (−1.32) | Deity (−0.88) | Gossip (−1.39) | Happiness (−0.89) | Contract (−0.79) |
| Gravity (−1.38) | Forgiveness (−0.99) | Necessity (−1.43) | Force (−1.30) | Faith (−0.93) |
| Acceleration (−1.56) | Pride (−1.19) | Anger (−1.45) | Heat (−1.55) | Equality (−1.18) |
| Sadness (−1.60) | Happiness (−1.38) | Sacrilege (−2.12) | Gravity (−2.07) | Force (−1.46) |
| Anger (−1.76) | Sadness (−1.42) | Compliment (−2.55) | Spirituality (−2.60) | Sacrilege (−1.61) |

## Combined-language factor analysis

The commonality of neural representation for abstract concepts in the two languages was characterized by six underlying dimensions, five of them readily interpretable. Of the five interpretable dimensions, four were semantic in nature, which we have labeled: Verbal representation, internality–externality to self, rule-based, and social content. (Each dimension is further described in the Discussion). The remaining non-semantic dimension corresponded to the length of the printed word that named the concept. The five interpretable group-level factors accounted for 36% of the variance in the participant-level factors. All but one of these factors (rule-based concepts) have been identified in a previous study of abstract concepts (Vargas & Just, 2020). The brain locations of the voxel clusters with high loadings on the interpretable factors are shown

as spheres in Figure 3.2. The concepts located at the extremes of each of these dimensions and their respective factor scores are shown in Table 3.4.

Behavioral ratings of each concept reflecting the saliency of each dimension (as it had been interpreted) were used as independent variables in a linear regression model that predicted the activation level of each concept in the factor locations. The mean rank accuracy of predictions for left-out concepts, averaged first over concepts and then over participants, was 0.73, $p < 0.001$. Performing the predictive modeling analysis while excluding the word length dimension resulted in a mean classification accuracy of 0.72, $p < 0.001$, which was not significantly different from the accuracy when word length was included, $t(27) = 1.80$, n.s.

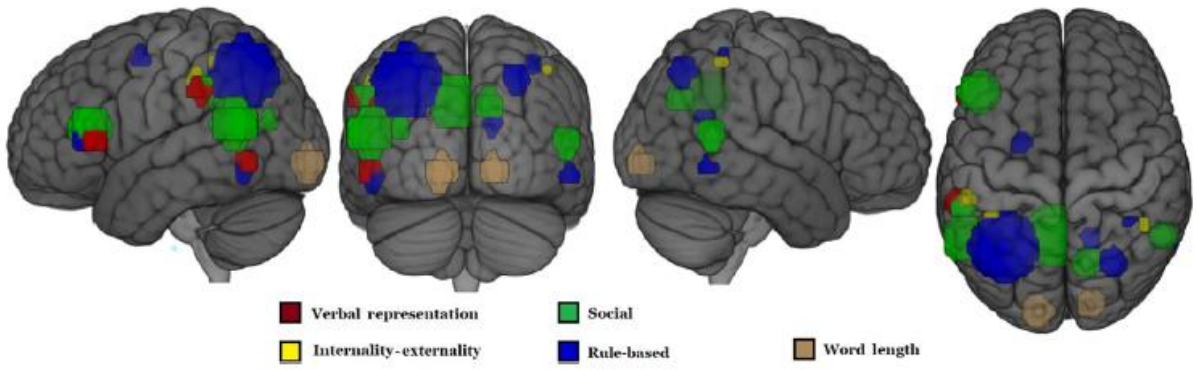


**Figure 3.2.** Locations for five interpretable factor dimensions from combined-language analysis. These spheres were specified using the centroids of clusters of voxels (containing a minimum of 10 voxels) with high loadings (>0.4) on each of the factors.

The correlation for a given dimension between the factor scores of the 28 concepts and their behavioral ratings were reliable for all semantic dimensions for both languages. The correlations between ratings and factor scores from the language-specific factor analyses for each semantic dimension are as follows: the externality dimension had an $r = 0.63$, $p < 0.001$ for Mandarin and $r = 0.70$, $p < 0.001$ for English; the social dimension had an $r = 0.52$, $p < 0.01$ for Mandarin and $r = 0.46$, $p < 0.05$ for English; and the rule-based dimension had an $r = 0.40$, $p < 0.05$ for Mandarin and $r = 0.39$, $p < 0.05$ for English. The similarity between languages in the correlations between factor scores and mean behavioral ratings for the Verbal dimension: in the case of the English ratings, it was $r = 0.82$, $p < 0.001$, and for the Mandarin ratings, it was $r = 0.42$, $p < 0.05$. These significant correlations between the behavioral ratings and factor scores indicate convergent validity for the interpretations of the semantic dimension for both English and Mandarin samples.

**Table 3.4.** The six concepts with the highest and lowest factor scores for each interpretable dimension from the combined-language factor analysis.

| Verbal representation | Externality/internality | Rule-based | Social content | Word length |
|---|---|---|---|---|
| Ethics (1.23) | Causality (1.79) | Multiplication (2.00) | Intimidation (1.40) | Causality (1.88) |
| Spirituality (1.19) | Gravity (1.53) | Subtraction (1.97) | Pride (1.33) | Acceleration (1.73) |
| Faith (1.19) | Sacrilege (1.29) | Probability (1.61) | Gossip (1.29) | Happiness (1.43) |
| Sacrilege (1.06) | Equality (1.06) | Acceleration (0.85) | Forgiveness (1.21) | Probability (1.19) |
| Necessity (0.89) | Subtraction (0.99) | Ethics (0.82) | Exoneration (1.11) | Compliment (0.99) |
| Exoneration (0.82) | Crime (0.67) | Contract (0.74) | Anger (0.72) | Necessity (0.86) |
| Sadness (−0.73) | Anger (−0.90) | Crime (−0.77) | Subtraction (−0.63) | Truth (−0.90) |
| Happiness (−1.19) | Forgiveness (−0.93) | Exoneration (−0.93) | Happiness (−0.65) | Ethics (−1.05) |
| Acceleration (−1.67) | Pride (−1.36) | Consciousness (−1.08) | Necessity (−0.77) | Pride (−1.24) |
| Force (−1.73) | Spirituality (−1.47) | Sacrilege (−1.30) | Multiplication (−0.92) | Faith (−1.52) |
| Gravity (−2.06) | Happiness (−1.82) | Anger (−1.36) | Heat (−1.77) | Crime (−1.59) |
| Heat (−2.12) | Sadness (−2.06) | Compliment (−1.70) | Spirituality (−3.15) | Force (−2.33) |

Common representation supported by cross-language classification

To assess the similarity of individual concept representations across languages based on the underlying factors, a classifier was trained on one language to predict individual concepts in the other language. Cross-language classification of individual concepts resulted in a mean rank accuracy (averaged over concepts, direction of decoding, and participants) of 0.65, $p < 0.001$. (The right-most column in Table 3.2 shows the accuracies for individual concepts.) When individual concepts were averaged within categories, the categories with the highest accuracies were **mathematics** (0.67), **scientific** (0.78), **emotion** (0.70), and **social** (0.70). When the features for the between participant cross-language classification were defined independently of the factors, using the union of 120 stable voxels from each language regardless of their association with any of the factors, the mean accuracy was 0.69, $p < 0.01$. The

classification accuracy was only slightly lower (0.65 vs. 0.69) when computed using only stable voxels associated with factors, indicating how well a factor-based account of the data accounts for the similarity between languages in their neural representations of abstract concepts.

*Differences in the neural representation of concepts across languages*

Although a common set of dimensions was identified (as indicated by the reliable correlations of factor scores across the two language specific factor analyses in Table 3.5), the distribution of the items along corresponding factors was similar but not identical (Table 3.3 and Table 3.1). Independently collected behavioral ratings provided converging evidence that a few individual abstract concepts are represented somewhat differently along the verbal representation dimension. English speakers rated emotions (e.g., *happiness* and *anger*), social concepts (e.g., *intimidation* and *compliment*) and spiritual concepts (e.g., *deity* and *sacrilege*) as being more verbally represented than did Mandarin speakers. Additionally, Mandarin speakers rated mathematical concepts (e.g., *subtraction* and *multiplication*) and scientific concepts (e.g., *heat* and *acceleration*) as more verbally represented than did English speakers.

**Table 3.5.** Correlation matrix of factor scores across English (rows) and Mandarin (columns) factor analyses for each semantic dimension.

| English-by-Mandarin | Verbal representation | Word length | Externality/internality | Rule-based | Social content |
|---|---|---|---|---|---|
| Verbal representation | 0.55 ($p < .01$) | −0.05 | −0.29 | −0.38 | 0.31 |
| Word length | 0.01 | 0.20 (n.s) | −0.33 | −0.11 | 0.14 |
| Externality/internality | 0.42 | 0.08 | 0.32 ($p < .1$) | −0.14 | 0.13 |
| Rule-based | 0.12 | 0.08 | 0.35 | 0.45 ($p < .05$) | 0.40 |
| Social content | −0.42 | −0.23 | −0.28 | 0.05 | 0.42 ($p < .05$) |

*Note*: The variance each dimension accounted for varied across languages but were aligned here for easier comparison.
Abbreviation: n.s, not significant

Qualitative descriptions of the concept properties that participants reported suggest polysemous words such as equality were represented somewhat differently across languages. English speakers tended to interpret the concept of equality partly in the context of social equality while Mandarin speakers tended to interpret equality in terms of its mathematical meaning. These results suggest that the difference between languages is not in brain function but in the meanings of the "translation equivalents" of the polysemous word equality in the two languages. These differences could be due in part to differences in relative frequency or prominence of the two senses of the word in the two languages.

## Discussion

### *Overview*

The results of this study suggest that a common neural infrastructure exists for representing abstract concepts across English and Mandarin. Factor analyses using activation data from both languages revealed four semantically interpretable dimensions, **verbal representation**, **internality– externality**, **social content**, and **rule-based** representation underlying the activation patterns of 28 abstract concepts for both languages. A secondary finding was that although the neural regions or systems associated with these representations are common, the representations of individual concepts sometimes differ with respect to the salience of an underlying dimension.

### *Language-invariant semantic primitives of abstract concept representation*

The four emerging neurosemantic dimensions underlying representation of abstract concepts are: **verbal representation**, **internality–externality**, **social content**, and **rule-based** representation. The ability to think of abstract meaning draws on the ability to think of concepts in terms of other verbal concepts, to rely on the use of the self as a reference, to think in terms of social contexts, and to consider the

rules that govern certain concepts. The combined-language factors highlight possible subnetworks that are present in the general semantic network identified by Ralph et al. (2017). These subnetworks represent the processing of specific types of semantic information. Regardless of language, people rely on the same broad neural systems to represent abstract concept meaning.

Although previous research has found emotion processing to be involved in the processing of abstract concepts (Kassam et al., 2013; Kousta et al., 2011; Vigliocco et al., 2014), an emotion-specific dimension did not emerge in our factor analyses. However, one of the factor locations of the social content dimension identified from the combined-language factor analysis, posterior cingulate, is often associated with affective processing. The inclusion of this region could be due to the affective involvement in abstract concepts like intimidation, pride, and forgiveness for the social dimension. Vigliocco et al.'s (2014) abstract stimuli were selected on a different basis (concreteness ratings) than ours (membership in seven abstract semantic categories). Many of our abstract categories (such as science, mathematics, and metaphysics (logic) and hence their members (e.g., gravity, force, acceleration, and causality, truth, necessity) are sparsely represented in the Vigliocco stimulus set. By contrast, the Vigliocco stimulus set contains many descriptors of mental states that have a significant affective component, such as agony delirium, frenzy, and panic. It may be that an affective

component plays a larger role in the representations of abstract concept representations pertaining to mental states rather than to physical or constructed worlds.

The verbal representation dimension of abstract concepts

This dimension organizes concept representations based on their degree of association with word (verbal) representations (manifested as activation in left inferior frontal gyrus (LIFG, specifically, the triangular subregion) and disassociation with visuospatial processing and action imagery (lower activation in LLOC and LSMG) (Vargas & Just, 2020). The verbal representation dimension accounted for the most variance (9%) in the participant-level factor representations and was the most salient dimension for all participants in both languages. Concepts such as *faith*, *spirituality*, and *ethics* anchor the verbal extreme of this dimension while concepts such as *heat*, *gravity*, and *force* anchor the other, nonverbal extreme. The concepts at the verbal extreme evoke activation in language areas, presumably because the concept evokes the thought of verbal labels for other related concepts. The presence and salience of this dimension in these data and in previous studies work (Vargas & Just, 2020; Wang et al., 2010) provide converging evidence for the prominent role of this semantic dimension in the representation of abstract concepts.

The self-based dimension of abstract concepts

The internality–externality dimension organizes conceptual representations based on the degree to which a concept uses the self as a reference. Concepts such as *pride*, *spirituality*, and *sadness* anchor the internal extreme of the dimension while the concepts *causality*, *gravity*, and *sacrilege* anchor the external extreme. This dimension accounted for 8% of the variance in the participant-level factor analysis. This factor's locations include RSMG, a region shown to be related to the projection of one's own mental state onto others (Silani et al., 2013).

The social interaction dimension of abstract concepts

This dimension organizes abstract representations based on whether they entail a social component. The concepts *intimidation*, *pride*, *gossip*, and *forgiveness* typify the extreme of this dimension. This dimension accounts for 6.4% of the variance in the participant-level factor analysis. The factor locations include regions associated with autobiographical information processing (posterior cingulate) and theory-of-mind (right temporoparietal junction), as well as the triangular region of LIFG. Although the triangular region of LIFG was a component of the social dimension in the combined-

language factor analysis, the presence of this component emanates from the Mandarin-specific data (Figure 3.1). This participation of LIFG in the Social dimension only in Mandarin is an example of differential involvement of various components of meaning across languages. It is uncertain from our study what functional role LIFG is contributing to this dimension; for example, it is possible that some abstract concepts with a social component also entail an associated verbal expression that LIFG references and this LIFG role may be differentially used in Mandarin.

The rule-based dimension of abstract concepts

This dimension organizes concepts in terms of their being based on some set of rules or that define or are defined by specific, precise relationships between other concepts. The concepts *multiplication*, *probability*, and *ethics* typify this dimension. The factor locations are left precuneus and right supramarginal gyrus. This factor accounts for 7.1% of the variance in the participant-level factor analysis. The regions associated with this dimension share a partial overlap in left parietal cortex with regions previously identified to be associated with algebraic and equation-based processing (Mason & Just, 2016). The Mandarin-specific factor analysis indicated that one of the rule-based factor locations was in the triangular region of LIFG (Figure 3.1). This region

is sometimes associated with semantic selection (Badre et al., 2005), and its role in the rule-based representations may entail reference to a verbal expression of some aspect of a rule.

The neural representation of word length in English and Mandarin

This dimension organizes the concepts in terms of the length of the written word that names them. For English, the word length factor scores are correlated with the number of characters in a word, with r = 0.68, $p < 0.001$. For Mandarin concepts, the number of strokes was used as a measure of word length. There was insufficient variance in the number of characters of the set of Mandarin concepts, with 24 of the 28 concepts containing two characters (M = 2.04; SD = 0.5). For Mandarin, the word length factor scores are correlated with the number of strokes in a word, with r = 0.56, $p < 0.01$. Although the number of characters and the number of strokes in a word are not correlated (r = 0.2, n.s.), when normalized (converted to z-scores) and averaged, the resulting relative word-length measure of participants of both languages was correlated with the factor scores of this dimension at r = 0.8, $p < 0.001$. The region associated with this dimension (i.e., where the voxels with high loadings on this factor are located) is

isolated to the occipital pole. This finding suggests the presence of a language-common, word length dimension that captures some aspect of the early visual percept of the word.

*Commonality of meaning for abstract concepts*

The common neural organization of the concepts, as characterized by the shared underlying semantic dimensions, was sufficient to serve as a basis for reliably classifying (identifying) individual concepts using the activation signature from the other language. This cross-language classifiability provides converging evidence for a language-invariant semantic representation of abstract concepts across the set of 28 concepts. The high cross-language decodability of mathematics and science concepts could be attributable to the language-invariant algebraic expression of these concepts (Mason & Just, 2016). The high cross-language decodability of emotion and social concepts could be attributable to a common embodied nature of these concepts.

The similarity of the factor analysis outcomes in the two languages indicates a common underlying neurocognitive infrastructure for processing abstract concepts. This semantic resource distributed across multiple brain locations constitutes a more specialized subset of regions previously identified as a general semantic network (Ralph et al., 2017). These regions were sufficient for reliably decoding individual abstract

concepts across languages based on their neural representations. However, some of the underlying dimensions were differentially salient across languages.

*Nuances in the meaning of individual abstract concepts across languages*

Not all concepts were decoded equally accurately across languages; in each language, there were individual concepts that were reliably classified within language but not across language. For example, *causality* was highly decodable within English (0.89 accuracy) and Mandarin (0.8) but less decodable across languages (0.62). In these cases, the semantic properties (e.g., contexts or associated concepts) that were generated by participants for these concept representations could be homogenous within each language but distinct across languages.

Low cross language decodability could often be explained by a differential salience of verbal processing (as is this case for some concepts such as *consciousness*, *necessity*, or *faith*) while other concept differences could be attributable to differing senses of meaning across languages. One such example cited above is that in Mandarin, the concept, *equality*, was more strongly interpreted in its mathematical sense than social sense relative to the English interpretation. These differences could be caused by differences in how concepts like equality are learned or taught, or the differences may

arise because of the polysemy of some of the words used to describe the abstract concept. Due to small sample sizes, it is uncertain whether cultural differences or characteristics of the two samples of participants or nonequivalence of the stimulus words' connotation or senses were responsible for the differences between languages in some of the neural representations.

In addition to occasional nuances of a difference between languages in the neural representations of individual concepts, the underlying dimensions sometimes played a larger role in one language than the other. For example, the verbal representation dimension accounted for 10% of variance in the English factor analysis but only 7% in the Mandarin analysis. Differences such as these indicate that abstract concepts in the two languages can evoke various dimensions of meaning to different degrees. Such differences could be attributable to differences in the senses or connotation of the translation-equivalent word concepts.

## Conclusion

Factor analyses revealed a set of common neurosemantic dimensions that constitute the basis for the representation of abstract concepts across languages: verbal representation, internality–externality, social content, and rule-based content. The

subsequent predictive modeling based on behavioral ratings of the concepts provides convergent validity for the factor interpretations. The successful cross-language classification suggests that the underlying semantic dimensions provide a sufficient basis for decoding abstract word concepts across languages. Although the neural dimensions used for representing abstract concepts are common across languages, differences in the meaning of some individual concepts can be accommodated in terms of differential salience of particular dimensions. These semantic dimensions constitute a set of neurocognitive resources for abstract concept representation within a larger set of regions responsible for general semantic processing.

# Chapter 4. Project 3: Representational differences in Socioenvironmental Concepts

The results of the study described in Chapter 3 suggest there is a common semantic structure underlying abstract concepts across linguistically defined groups. However, there is little reason to expect large differences in the meaning of the abstract concepts or their associations across languages. That is, across languages, the definition of concepts' meaning is not hypothesized to differ, nor are they expected to differ in cultural contexts or situations they are experienced. For example, a concept such as *contract* refers to a legally binding agreement between two persons or entities regardless of culture or language. There are no systemic or historical incidents that would bias the semantic association for native speakers of either language just as the concept of *multiplication* is likely experienced in classrooms or in the context of engineering problems for both Chinese and American cultures.

Across people that share a common language, however, there may be subtle variation in concept associations that are driven by differences in experiences. For example, marginalized communities often have largely different life experiences than individuals from dominant cultural groups. Thus, we would expect variability in concept associations based on such group differences within a larger, linguistically defined culture.

Chapter 4 attempts to address this question by contextualizing this it in an existing theoretical framework for describing the influence of environmental factors on concept associations, the Bias of Crowds model (Payne et al., 2017). This model asserts that environmental factors (i.e., contexts and situations) are reflected in concept associations. Although much of this research has examined the associations of attributes and concept categories towards specific racial groups, this framework could be used to describe how racial disparities in the United States (particularly between White and Black Americans) influence how people associate institutions where these disparities are present. These institutions belong to a family of concepts we will refer to as socioenvironmental concepts.

## Introduction

The concept *apple* refers to a round sweet fruit regardless of whether you say it in English or Mandarin just as a *contract* refers to a binding agreement. There is evidence for a common neural basis (as measured using functional Magnetic Resonance Imaging (fMRI)) in the semantic organization of abstract concept representations across languages (Vargas and Just, 2022). Despite this commonality there is little reason to hypothesize differences in the meaning of concepts beyond words having polysemous meaning in one language and not the other. However, research on the implicit biases of

racial groups provides a framework for hypothesizing differences in the semantic organization of a set of concepts.

Behavioral research on implicit concept associations, contextualized as the Bias of Crowds model, suggests that the associations between concepts reflect environmental factors (broadly described as contexts or situations) (Brown-Iannuzzi et al., 2017; Lee et al., 2017; Payne et al., 2017; Vuletich and Payne, 2019). Contexts can be as proximal as a specific personal encounter or as distal as the systemic and cultural norms of the nation the individual resides in. Contexts can also be as isolated as a one-time incident or as chronic as repeated exposure to associations such as media exposure. Although much of this research has focused on examining implicit biases of racial groups, the Bias of Crowds framework could effectively be used to explain the relationship between racial disparities (specifically in the context of Black and White Americans) and differences in the semantic organization of societal institutions where those disparities are present. A proper examination of systematic differences in conceptual associations across racial groups would require sample sizes that are practically outside the scope of traditional functional fMRI studies. The prohibitive expense of fMRI studies often limits sample sizes to 10-20 participants. Thus, to measure differences in conceptual associations, two gold-standard measures of concept geometry (Spatial Arrangement Method (SpAM) and the Pairwise Rating Method (PRaM)) will be utilized in addition to a novel proposed

implicit measure of concept geometry, the Implicit Semantic Association Procedure (ISAP). The first study of this chapter assessed the reliability and validity of these three metrics using a set of simple object concepts and attributes with clearly defined associations. The second half of this chapter assessed whether racial disparities between Black and White Americans are reflected in the association of concepts related to institutions within our American society (e.g., *healthcare, police*).

## Background

*How do our environments influence how we think?*

The Bias of Crowds framework asserts that implicit associations between concepts reflect environmental factors (broadly described as contexts or situations) (Payne et al., 2017). Contexts can be as proximal as a situational personal encounter or as distal as the systemic and cultural norms of the nation the individual resides in. Contexts can also be as isolated as a one-time incident or as chronic as repeated exposure to associations. Although measures of implicit association within an individual have been shown to be transient across time (Payne et al., 2017), when averaged across individuals, individual-specific idiosyncratic noise in concept associations are eliminated and the influence of shared environmental and contextual influences become more salient (Vuletich and Payne, 2019).

This idea of averaging across individuals to gain a group-level understanding is reinforced by the wisdom-of crowds phenomena, which posits that partial knowledge distributed among many individuals culminates in a true reflection of a shared experience or shared set of concept associations (Surowiecki, 2004). This means that, although a concept could have the same definition to two individuals, its conceptual associations are constantly evolving. Concepts become more than a simple representation of an external world property when they are chronically associated with certain contexts or other concepts. Contexts such as the geographical state that we live in, identities that we affiliate with, media we are exposed to, and situational factors like mood influence the likelihood of associations between certain concepts. For example, the concept of *firefighter* and *police officer* share common features on a superficial definitional level (e.g., public servant) however they differ in how they are contextualized (e.g., media portrayal) thus modulating their conceptual associations. Moreover, when contexts differ along a group of people (as is the case with racial disparities), then differing associations should be reflected in the concepts linked to those disparities. The Bias of crowds and, by extension, the wisdom-of-crowds frameworks provide a schema for thinking about racial disparities between Black and White Americans in the United States and how these disparities could be reflected in the associations of socioenvironmental concepts central to those disparities.

*Racial disparities in the United States for Black and White Americans*

People of color in the U.S. have historically been disproportionately negatively affected by health, economic, environmental, and political disparities when compared with White/European- Americans. These differences in adverse outcomes and positive opportunities across racial identities are defined as racial disparities (Kendi, 2016). People of color die sooner (Umberson et al., 2017), are at greater risk of being killed by police use of force (Plant and Peruche, 2005; Edwards et al., 2019), are more likely to live in poorer neighborhoods (Firebaugh and Acciai, 2016), are more likely to be exposed to toxins in the work place (Ash and Boyce, 2018), are at greater risk of dying from cardiovascular disease (Mozaffarian et al., 2015), receive differential treatment by clinician (Ryn et al., 2011), and, most recently, have a higher chance of contracting and dying from COVID-19 (Kim and Bostwick, 2020; Millet et al., 2020; Tai et al., 2020). Taken together, the available evidence converges on the conclusion that racial and ethnic minorities in America navigate greater risks and dangers than those experienced by White Americans.

Polling research has provided some preliminary insight into the influence of these racialized differences on perceptions of societal infrastructure in America. Black Americans reported negative views towards police with 30% reporting very little to no

confidence in the police as compared to 7% of White Americans (Schuck and Rosenbaum, 2005). Of Black Americans, 63% reported that discrimination greatly hurts their chances of getting a good paying job while only 5% say it has little to no effect (Gay, 2004). Based on national polling data by the Gallup Poll Social Series from 2001 to 2015, Hispanic people and Black Americans are more concerned, relative to White Americans, about issues associated with air pollution, drinking water, and soil contamination (Lazri and Konisky, 2019). An examination of survey data from the General Social Survey database from years 2016 and 2018 revealed that when asked if "Black Americans have worse jobs, income, and housing than White people due to discrimination", 65% of Black participants agreed while only 40% of White participants agreed, $\chi^2$ (4, N = 2958) = 120.36, p < 0.0001 (Smith et al., 2019). These polling results in conjunction with empirical work on racial disparities provide converging evidence that the mental relations between social and environmental concepts may differ across racialized groups. Measuring whether racial disparities are reflected in the organization of institutions in our societal environment will require a measure of representational concept association.

*Measuring Representation Association*

Assessing representational geometry: SpAM and PRaM

Understanding the semantic organization of a set of concepts involves the examination of its inter-concept geometric structure (i.e., the collective set of associations). For example, Koch and colleagues (2020) had participants spatially organize a set of influential leaders on a 2-dimensional plane based on the (dis)similarity between leaders. Using these data across participants the authors were able to generate semantic organization principles participants used to arrange each stimulus (Figure 4.1).

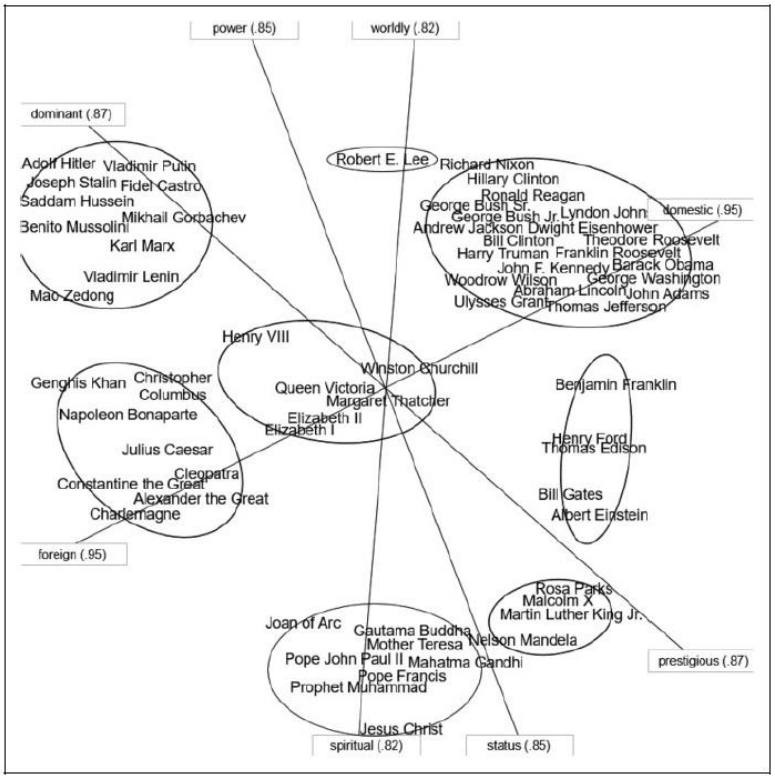


**Figure 4.1.** Arrangement of influential leaders and the underlying organizing dimensions as seen in Koch et al. (2020).

The gold-standard for measuring pair-wise associations and representational geometries the Pairwise Rating Measure (PRaM). This measure involved obtaining Likert-ratings of concept similarities for all unique combinations of concepts in a set and to apply multidimensional scaling to these pair-wise ratings (Verheyen et al., 2016). Despite the PRaM's clarity and reliability, it's main criticism is that the duration of the task exponentially increases with the number of concepts included in a set.

The Spatial Association Measure (SpAM) was designed to overcome the exponential increase in session duration. In a conventional SpAM task, participants spatially organize concepts on a screen such that the distance between stimuli represents their perceived (dis)similarity with shorter distances reflecting greater semantic association and longer distances reflecting associated dissimilarity (Goldstone, 1994). Participants provide their responses by dragging and dropping "more similar" concepts closer together and "more dissimilar" targets further apart on a 2-dimensional surface. Moreover, by organizing concepts as a set according to their similarity, participants can communicate more nuanced semantic properties (Verheyen et al., 2020). The SpAM measure, despite being more scalable with respect to stimulus set size, has been shown to favor spatial over featural representations and tends to limit the number of underlying dimensions that can emerge because the 2-dimensional nature of

the task biases participants to think in 2-dimensional terms (Verheyen et al., 2016). For visual and verbal stimuli, the SpAM has been shown to correlate strongly with the PRaM (Hout et al., 2013). However, the results obtained from the SpAM and the PRaM often converge considering that they are both measuring explicit concept association and are vulnerable to similar response biases. Implicit measures of concept association have been used to circumvent the possibility of response bias.

Implicit measures of concept association

One of the first implicit measures, the Implicit Association Task (IAT; Greenwald et al., 1998) has historically been used to measure implicit biased attitudes towards racial and ethnic groups. Using reaction time as a critical measure, participants supraliminally categorize exemplars belonging to a category (e.g., *Asian/European faces*) along a different dichotomous category (e.g., *Foreign/American*). More recently, the Affect Misattribution Procedure (AMP; Payne and Lundberg, 2014) has been used to measure biased attitudes by implicitly measuring Likert ratings of an irrelevant target stimuli which reflect misattributed attitudes towards a primed stimulus. The structure of AMP involves briefly presenting a prime stimulus (e.g., picture of an infant) followed by an ambiguous pictograph (e.g., chinese character); participants are then required to make a judgment of the pictograph (*Pleasant/Unpleasant?*). The response towards the target

(Chinese character) is indirectly influenced by the participant's attitude towards the prime (infant picture). The AMPs has been shown to predict alcohol consumption (Hofmann et al., 2010) and moral decisions (Hofmann and Baumert, 2010). Despite the IAT and AMPs successful ability to measure biased attitudes towards racial groups, the two measures are designed to measure implicit associations with respect to categories rather than all pair-wise associations of stimuli. Additionally, the design of the original AMP is parameterized to examine visually presentable stimuli.

Part of this research plan capitalizes on the supraliminal capabilities of implicit measures such as the IAT and AMP and holistic geometric information from the SpAM and PRaM by proposing a new methodology based on a modified version of the AMP paradigm designed to assess the semantic similarity/dissimilarity of any number of pairs of concepts by measuring differences in reaction time, the Implicit Semantic Accessibility Procedure (ISAP). Participants will be presented two words in quick succession (a prime and target) and then asked to answer a semantically irrelevant question ("*Is the target word English?*") (See Figure 4.2 for example trial of the ISAP). Shorter response times to the question (relative to a global baseline) are hypothesized to correspond to increased semantic similarity in the prime/target pairs while longer response times are hypothesized to correspond to semantic conflict present in the prime/target pairs. This difference in reaction time is represented as a differential in semantic accessibility – that

is, when the prime concept is cued, what is the probability that a particular target word would follow (Payne et al., 2017). These modifications allow for the comparison of all pair-wise relationships between stimuli rather than along a single dimension (as designed in the IAT and AMP) while circumventing possible response bias present in explicit measures concept geometry (i.e., SpAM and PRaM).

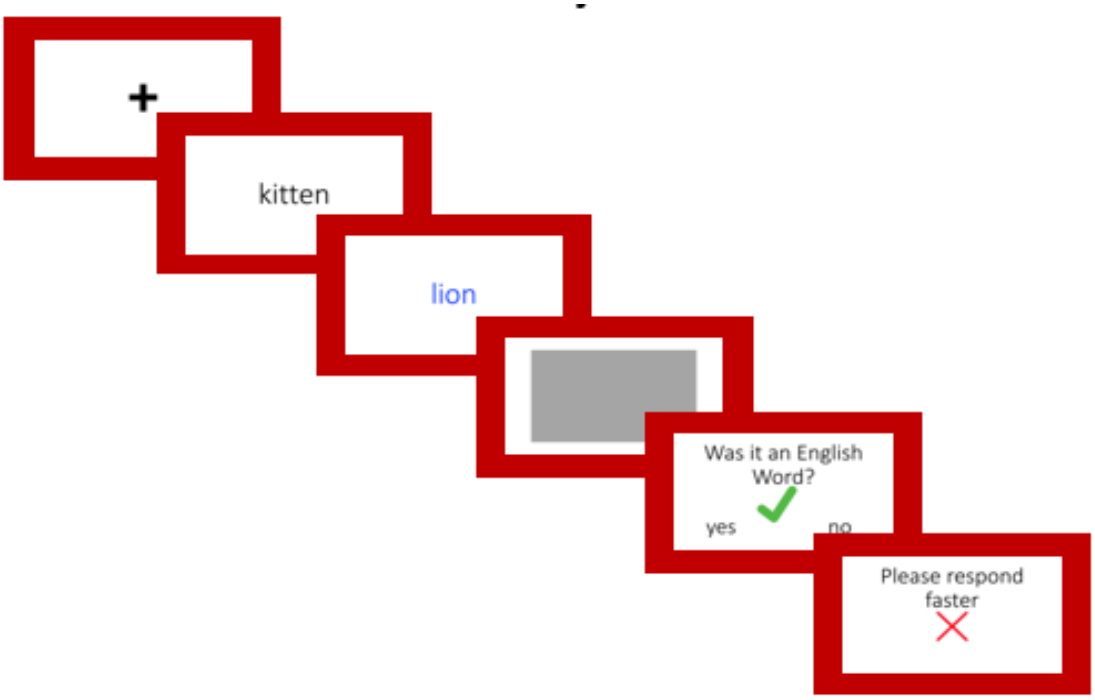


**Figure 4.2.** A sequence diagram of a ISAP trial. Each trial begins with a fixation cross. A pair of words (prime and target) are then presented in quick succession and are followed by a grey noise mask. Participants are then primed to provide a Yes or No response as to whether the target word (shown in blue) was a word in English or a non-word.

This project aims to measure whether racial disparities are reflected in the geometry of concept associations Black and White Americans have towards concepts

central to their societal environment. To measure differences in concept geometry the SpAM and PRaM will be utilized in addition to proposed ISAP. The first study of this chapter assesses the reliability and validity of these three metrics using a set of simple object concepts and attributes with clearly defined associations. The second study of this chapter assesses whether racial disparities between Black and White Americans are reflected in the association of concepts related to institutions within our American society (e.g., *healthcare, police*).

## Study 1: Assessing the Reliability and Validity of the ISAP, SpAM, and PRaM

In Study 1 the reliability and validity of semantic association measures (ISAP, SpAM, and PRaM) are assessed. The reliability is assessed internally for all tasks using a split-half measure (Hout et al., 2013; Verheyen et al., 2016; Verheyen et al., 2020) and the test-retest reliability is assessed specifically for the ISAP by correlating the pair-wise similarity 1-day apart. After screening for data-quality and incomplete sessions 75 participants completed all 3 tasks using a set of 8 concepts. The primary goal of this study is to determine the efficacy of the ISAP as a tool for measuring the semantic geometry (i.e., the pair-wise semantic association of a collective set of concepts) and to replicate past reliability statistics for the SpAM and PRaM.

## Methods

### *Participants*

One-hundred participants were initially recruited via Mechanical Turk using the Cloud Research toolbox (formerly TurkPrime; Litman et al., 2017) from a nationwide sample. ISAP, SpAM, and PrAM tasks were completed using Gorilla, a platform for implementing and hosting experimental paradigms for online testing (Anwyl-Irvine et al., 2020). These 100 participants were solicited to participate in the same ISAP task 1-day after their completion of the first session. Ten participants were excluded due to incomplete session and an additional 9 were excluded due to insufficient accuracy on attention checks during reaction-time measures of the ISAP task (< 90% accuracy on non-scrambled trials; <75% accuracy on scrambled trials). Insufficient accuracy was caused by a consistent lack of responding during trials or fixed responding (i.e., providing the same response for all questions regardless of scrambled/non-scrambled trials) during the ISAP task. The study sessions were automatically terminated if the total duration of the session exceeded 20-minutes past the expected study duration. Of the remaining 81 participants, 75 participants (93%) successfully completed testing on the second day. The final analyzed data set consisted of these 75 participants (6.67% Black; 76% White; 5.33% Asian; 12% Hispanic or Latino; mean Age = 34.33, SD = 6.97) collected from 25 different states with each participant residing in a different zip code.

*Stimulus set*

Stimuli were 8 concepts distributed among 3 categories. Although the categories were not presented to participants they are listed here for expository purposes preceding the list of concept exemplars: **(pets)**: *puppy*, *kitten*, *spider*; **(objects)**: *knife*, *gun*, *blanket*; **(attributions)**: *danger*, *soft*. Two of the categories were designed to correspond to concrete concepts while the final category corresponds to two attributions each of which are expected to align semantically with 2 concepts from one category and 1 from the other. Past neuroimaging work has shown that focusing on the representations of individual concepts provides a higher resolution of semantic content than examination on a categorical level (Just et al., 2010).

*Experimental Design*

*ISAP*

During the ISAP task each stimulus was presented 18 blocks to enable averaging of noise present in individual trials. The ISAP was structures similarly to the AMPs -- each trial consisted of a fixation (700ms), followed by a prime concept (225 ms), target concept (225 ms), a mask (150ms ms), and a question (1250 ms window of time to

provide a response) (See Figure 4.2 for a diagram of a single trial). Participants were shows two consecutive concepts and then asked whether the second concept was in English. The response time to the question is used as a proxy of semantic association between the prime and the target concept. For the set of 8 concepts, there were a total of 8 choose 2 (i.e., 28) unique concept combinations. In each block all possible pairs of stimuli were shown. For each block, additional scrambled trials were included (proportional to 25% of the unique combinations). In these scrambled trials, the target concept was a scrambled version of a concept from the set. For example, if *tiger* was a candidate target concept than a possible scramble could be *rgite*. These scrambled trials served primarily as an attention check. The reaction time was the critical measure used for analyses. Half-way through the ISAP task the left-right input responses were inverted. This was done to accommodate handedness bias that may be present and to be consistent with designs of modern versions of the IAT (https://www.projectimplicit.net/).

*Pairwise-rating of Association Measure (PRaM)*

During the PRaM task participants provided explicit Likert ratings (1 to 7) representing semantically (dis)similar for a pair of concepts. A rating of 1 indicated that the pair of shown concepts are extremely dissimilar while a rating of 7 indicated that the

pair of concepts were extremely similar. Participants provide rating for all 28 possible combinations of the 8 concepts. Only a single set of rating was obtained from participants.

*Spatial Arrangement measure (SPaM)*

During the SPaM task, participants were instructed to spatially arranged boxes labeled with each of the 8 concept according to their semantic similarity. Participants were instructed to arrange concepts that were semantically similar are closer together while concepts that were spatially dissimilar were farther apart. The X and Y coordinates of the pixels for a concept's placement within the display were used to determine the spatial arrangement of each concept. These XY coordinated were used to compute Euclidean distances between all combinations of concept pairs.

*Data processing*

For the ISAP, reaction times for all 28 unique pairs of the 8 concepts were averaged across all 18 blocks to generate a stable estimate. Timed-out trials were excluded from the mean. Mean reaction times were then normalized within each participant to control for participant-specific differences in mean reaction times. For

each pair of concepts positive z-scores reflect slower reaction times relative to all the pair-wise relationships between concepts and negative z-scores reflect pairs of concepts where the participant responding relatively quickly. To make the polarity of semantic associations consistent across tasks the vector of 28 unique concept pairings was multiplied by -1 to invert the polarity on some tasks such that all positive z scores reflected greater semantic similarity and negative z scores reflected less semantic similarity. Pair-wise associations for the SpAM and PRaM were similarly normalized within each participant. Z-scores for the SpAM were likewise multiplied by -1 to similarly invert the polarity on some tasks such that all positive z scores reflected greater semantic similarity and negative z scores reflected less semantic similarity. For each of the 3 tasks, these normalized 28 pairwise associations were used for subsequent analyses.

*Assessing internal consistency and test-retest reliability*

For the SpAM and PRaM tasks, reliability was computed as the split-half correlation of the vector of 28 pair-wise similarities (Hout et al., 2013; Verheyen et al., 2016; Verheyen et al., 2020). The participant data were separated into two halves and the averaged normalized similarity vectors were correlated across halves within each

participant. To obtain a stable estimation of split-half reliability, this analysis was repeated 10,000 times randomizing the participant distribution between the two halves across permutations. To obtain this metric of relative reliability a factor k coefficient was computed, representing the multiplier to the number of the participants recruited required to obtain a desired reliability. Using the PRaM reliability as the gold-standard, a factor k was computed indicating the magnitude more participants that would need to be tested using SpAM or ISAP to attain the same level of reliability as the PRaM. Thus the factor k represents the multiplier to the number of the participants recruited required to obtain reliability desired for the PRaM (Equation 4.1; Lord and Novick, 1968):

**Equation 4.1.** Factor k coefficient equation for measuring relative reliability.

$$k = \frac{\rho_D(1 - \rho_O)}{\rho_O(1 - \rho_D)},$$

the desired reliability $\rho_D$ equal to PRaM's averaged split-half reliability and the observed reliability $\rho_O$ is equal to SpAM or ISAP's reliability.

Concept associations are expected to vary across time because contexts are expected to vary across time (Payne et al., 2017). This property of concepts often results in low test-retest reliability of individual participant has been shown to be low when

measured over long intervals of time. Therefore, for this study, the test-retest reliability was calculated across a relatively short window of time (1-Day) and across groups (Table 4.1).

*Assessing the validity of the ISAP*

Using the post-processed data, the vector of 28 pair-wise similarities for the ISAP was correlated with the same vectors from the SpAM and the PRaM. If the ISAP is a comparable measure of concept association to the gold-standard then the correlation between the ISAP and the SpAM and PRaM should be comparable to the correlation between the SpAM with the PRaM.

**Table 4.1.** Correlation matrix of the vectors of pair-wise semantic similarity of the ISAP (including both days), SpAM, and PRaM.

| Variable | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1. ISAP Day 1 | - | | | |
| 2. ISAP Day 2 | .46* | - | | |
| 3. SpAM | -.03 | -.20 | - | |
| 4. PRaM | -.11 | -.19 | .94*** | - |

Note: * indicates $p < 0.05$; *** indicates $p < 0.001$

## Results

*Split-half and test-retest reliability*

Table 4.2 lists the average split-half reliability averaged across iterations for each of the measures as well as factor k metrics relative to the gold-standard PRaM. Both SpAM and PRaM measures yielded high average reliabilities (0.96 and 0.99 respectively). By contrast, the ISAP yielded a reliability of 0.41. As a result, k coefficients reflected that more participants need to be tested using the SpAM (by a factor of 5.65) or the ISAP (by a factor of 224.02) to obtain theoretically comparable reliability to the PRaM. The test-retest reliability of the ISAP across 1 Day yielded a correlation of $r = 0.46$, $p < 0.05$.

**Table 4.2.** Averaged split-half reliability for 10,000 iterations and factor k coefficients. Factor k coefficients were computed with the PRaM being the desired target reliability. Factor k coefficients represent the multiplier to the number of participants recruited in the observed task required to obtain a comparable reliability to the desired task.

| Task | N | ρ | k |
|---|---|---|---|
| ISAP | 75 | 0.41 | 224.02 |
| SpAM | 75 | 0.96 | 5.65 |
| PRaM | 75 | 0.99 | |

*Validity of the ISAP*

Validity for the ISAP is assessed by the correlating the concept geometry of the 28 unique concept-pair combinations of the ISAP with both the SpAM and PRaM. Both days of the ISAP were shown to be non-significantly related to the other gold-standard measures of concept geometry (Table 4.1). By contrast, the SpAM and PRaM were shown to be highly correlated ($r = 0.94$, $p < 0.001$).

## Discussion

The SpAM and PRaM are both highly reliable measures of concept geometry while the ISAP showed to be relatively considerably less reliable. The reliability statistics obtain for the SpAM and PRaM are consistent with previously reported reliability measure using an identical metric for an identical number of exemplars (PRaM, rho = 0.98 and SpAM, rho = 0.94; Verheyen et al., 2020). Although the ISAP may not be a compartable to the gold-standard measures of concept geometry, it may still be an effective measure of implicit association.

### *Comparing the viability of the ISAP to other measures of implicit association*

Although the ISAP has been shown to not be as reliable as other gold-standard explicit measures of concept geometry it is potentially an effective measure of implicit association. The ISAP yielded considerably lower reliability as compared to another measure of implicit associations, the AMP. A meta-analysis of AMP suggest the task yielded Cronbach's alpha ranged from 0.47 to 0.95, with an average alpha of 0.81 (95% confidence interval= 0.77, 0.85). This meta-analysis did not include estimates of the test-retest reliability (Payne and Lundberg, 2014). There have not been any known direct measures of test-retest reliability for the SpAM and PRaM. However, the IAT,

another implicit association measure, has been shown to have a test-retest reliability of around 0.50 (Greenwald et al., 2021) which is roughly similar to the test-retest reliability obtained by the ISAP ($r = 0.46$, $p < 0.05$). This comparable reliability may implicitly suggest that the ISAP and IAT are measuring similar constructs (e.g., "gut reactions") or are measuring constructs with similar mechanistic properties (e.g., transient across time). If the ISAP and IAT are measuring similar constructs, then that would further cement the innovation of the ISAP as a tool for measuring conceptual geometries in a supraliminal manner. A comprehensive examination of the ISAP would require a direct comparison between the ISAP and IAT.

There is evidence to suggest implicit measures of concept associations based on "gut reactions" while explicit measures of concept association measure peoples' "considered options" (Payne et al., 2017). Moreover, participants were able to accurately predict their bias under circumstances where implicit measures of association were acquired (Ranganath et al., 2008). To assess whether the lack of correlation between the ISAP and the SpAM and PRaM reflects this distinction a follow-up study would be required. Theoretically the ISAP should reflect similar associations found with other implicit measures of concept association such as the IAT and the AMP. Considering that Study 1 utilized only 2 attributions (i.e., *soft* and *danger*), it would be possible to replicate the proposed study using an IAT paradigm. However, considering that the

ISAP was found to be distinct in what it is measuring relative to the SpAM and PRaM as evident by low cross correlations. Additionally, because of the low relatively low reliability of the ISAP, it is difficult to assess whether it is all-together measuring a distinct construct or whether there is an excess of noise in the signal. Taken together, the SpAM and PRaM have been shown to be effective measures of concept geometry and will be the primary measures of focus for Study 2.

## Study 2: Measuring Socioenvironmental concepts across Black and White Americans

The goal of the second study was to assess the representational geometry of a set of concepts that are hypothesized to differ in their conceptual associations across racial groups. Racial disparities for institutions (e.g., *healthcare, police)* within the United States between Black and White Americans contextualize the environments each racial group experience these institutions. These disparities serve a basis for identifying the influence of environmental factors on concept associations across racial groups. Using the tasks from Study 1, pair-wise semantic associations were measured across Black and White participants to identify possible differences in semantic geometry of socioenvironmental concepts.

## Methods

*Participants*

One-hundred participants (50 Black; 50 White) were recruited via Mechanical Turk using the Cloud Research toolbox (Litman et al., 2017). Six participants were excluded due to incomplete sessions. An additional 10 participants were excluded due to insufficient accuracy on attention checks during the ISAP task (< 90% accuracy on non-scrambled trials; <75% accuracy on scrambled trials). The final sample contains 84 participants (40 Black, mean age = 33.85, SD = 8.59, 93% Female; 44 White, mean age = 35.73, SD = 7.60, 77% Female). Participants were collected across 78 unique zip codes. The ISAP, SpAM, and PRaM were implemented using Gorilla, a platform for implementing and hosting experimental paradigms for online testing (Anwyl-Irvine et al., 2020).

*Stimulus set*

Stimuli were 7 concepts distributed among 2 categories: **(socioenvironmental concepts)**: *police*, *healthcare*, *media*, *community*, *voting*; **(attributions)**: *trust*, *fear*. The selection of these concepts was informed by empirical research on racial disparities

across White and Black Americans. To limit the session duration, the concepts chosen were not intended to be exhaustive of all aspects of societal infrastructure.

*Experimental Design*

Participants completed the ISAP, SpAM, and PRaM during a single 1-hour online session. Each participant completed 18 blocks of the ISAP procedure. The structure of individual trials for the ISAP was identical to the Study 1. For the 7 concepts, there were 21 unique combinations of stimuli, and all unique concept-pairs were presented in a single block. The structure of the SpAM and PRaM tasks were also identical to Study 1.

Following the completion of the three concept association tasks, a set of questionnaires were completed to be used to explore possible mechanisms of concept association: a short-form of the Social Dominance Orientation scale (Ho et al., 2015); explicit rating of engagement and valanced attitudes towards each of the 5 socioenvironmental concepts; self-reported perceptions of gender and racial discrimination; ratings of comfort providing opinions for each of the 5 socioenvironmental concepts to strangers. Additionally, participants who reported their race as Black or African American during the demographic intake were asked to complete the Centrality Scale of the Multidimensional Inventory of Black Identity (MIBI; Sellers et al., 1997).

*Data processing*

Similarly to Study 1, pair-wise associations for the ISAP, SpAM and PRaM were normalized within each participant. Z scores for the ISAP and SpAM were multiplied by -1 to similarly invert the polarity on some tasks such that all positive z scores reflected greater semantic similarity and negative z scores reflected less semantic similarity. For each of the 3 tasks, these normalized 21 pairwise associations were used for subsequent analysis.

*Reliability of semantic association within racial groups*

The reliability statistics were measured using an identical methodology as Study 1. Split-half reliability was measured for each racial group and task separately. The PRaM was used as the reference task for computing factor k (Equation 1).

*Measuring overall similarities and differences in representational structure by task*

To assess task-level similarities and differences across all 21 unique pairs of concepts data for each task and racial group were averaged across participants. The

task-specific vectors of 21 concept-pair associations were then correlated across task and group. These correlations provide a macroscopic view of the relationship across tasks within the same racial group and how the three tasks compare across groups.

*Measuring differences in specific concept pairs*

To gain a more nuanced understanding of how racial groups differ for individual socioenvironmental concept pairs, a permuted linear regression approach was utilized. Treating racial identity as a coded variable (Black = 1; White = 0), concept pair similarity was predicted. A separate regression model was created for each of the 21 concept pairs. Assessment of statistical significance was estimated using a 10,000-iteration permutation with racial identity being randomized across iterations. Age was entered as a covariate into each model and was held constant across permutations. Concept-pair relations were also qualitatively assessed using similarity matrices. Each element of the data structure reflects the normalized pairwise relationship averaged across participants. Separate data structures were computed for each racial group.

## Results

*Reliability of semantic association within groups*

The set of concepts in Study 2 generally yielded lower reliability than the set in Study 1 for all tasks. Despite being relatively lower, both the SpAM and PRaM had high reliability for each racial group (Table 4.3). The ISAP yielded low reliability for both racial groups and were comparable to Study 1. For both groups, the factor k suggests that 2.38 times (for the Black group) or 3.44 times (for the White group) more participants would need to be recruited for the SpAM to have comparable reliability to the PRaM.

Table 4.3 Split-half reliability of each task for each racial group

| ParticipantGroup | Task | N | ρ | k |
|---|---|---|---|---|
| Black | ISAP | 40 | 0.29 | 23.30 |
| | SpAM | 40 | 0.80 | 2.38 |
| | PRaM | 40 | 0.91 | |
| White | ISAP | 44 | 0.41 | 17.28 |
| | SpAM | 44 | 0.77 | 3.44 |
| | PRaM | 44 | 0.92 | |

*Measuring overall similarities and differences in representational structure by task*

Task-level similarities and differences across all 21 unique pairs of concepts data for each task and racial group inform of both cross-task consistency within each racial group as well as the relationships of similarity structures across racial groups. For both

groups, the ISAP was found to non-correlated with the SpAM and PRaM. However, there was a significant correlation between the SpAM and PRaM within each group (r = 0.49, $p < 0.05$ for Black participants; r = 0.66, $p < 0.01$ for White participants) (Table 4.4). Significant cross-group correlations were observed for both the SpAM (r = 0.77, $p < 0.001$) and PRaM (r = 0.92, $p < 0.001$). This indicates that there was correspondence in similarity structure across the set of 21 concept-pairs. However, despite this overall correspondence there may be specific concept pairs that differ across racial groups.

**Table 4.4.** Correlation matrix of the 21 socioenvironmental concept-pairs for ISAP, SpAM, and PRaM for Black and White Americans.

| Variables | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1. bISAPz | - | | | | | |
| 2. bSpAMz | -0.08 | - | | | | |
| 3. bPRaMz | 0.03 | 0.49* | - | | | |
| 4. wISAPz | 0.25 | 0.04 | 0.15 | - | | |
| 5. wSpAMz | 0.12 | 0.77*** | 0.54* | 0.22 | - | |
| 6. wPRaMz | 0.18 | 0.44* | 0.92*** | 0.24 | 0.66** | - |

Note: * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; Across group comparisions in grey

*Measuring Racial differences in specific concept pairs*

Racial group membership was used to predict pair-wise similarity for all 21 combinations between socioenvironmental concepts (Table 4.5). Given that Black was coded as 1 and White as 0, positive betas reflect instances were Black participants had greater similarity between concept pairs and negative betas reflect greater similarity among White participants. For the **PRaM**, *police-fear* was significantly more

semantically associated for Black participants and *police-trust* and *healthcare-trust* were significantly more associated for White participants. For the **SpAM**, *police-fear*, *healthcare-media*, and *media-voting* were significantly more associated for Black participants and *police-healthcare*, *police-trust*, and *community-voting* were significantly more associated for White participants. Lastly, for the **ISAP**, *police-media*, *police-trust* were significantly associated for Black participants and *police-community* and *healthcare-fear*, and *media-fear* were significantly associated for White participants. The differences identified by this analysis dissipated when corrected for False Discovery Rate. A qualitative assessment of concept pairs for each task and racial group can be seen in Figures 4.3-4.5.

**Table 4.5.** betas and p-values for predicting individual socioenvironmental concept pairs using racial identity as the predictor. Positive betas reflect greater similarity for concept pairs in Black participants while negative betas reflect greater similarity in White participants.

| Concept pair | | ISAP | | SpAM | | PRaM | |
|---|---|---|---|---|---|---|---|
| | | **beta** | **p** | **beta** | **p** | **beta** | **p** |
| police | healthcare | 0.14 | 0.228 | **-0.50** | **0.032** | -0.23 | 0.120 |
| police | media | **0.38** | **0.036** | 0.00 | 0.416 | 0.17 | 0.122 |
| police | community | **-0.45** | **0.018** | -0.34 | 0.101 | 0.09 | 0.410 |
| police | voting | 0.00 | 0.470 | 0.00 | 0.415 | 0.12 | 0.179 |
| police | trust | **0.56** | **0.006** | **-0.53** | **0.022** | **-0.44** | **0.026** |
| police | fear | 0.23 | 0.114 | **0.55** | **0.040** | **0.60** | **0.014** |
| healthcare | media | -0.04 | 0.418 | **0.45** | **0.008** | -0.05 | 0.455 |
| healthcare | community | 0.00 | 0.482 | -0.11 | 0.232 | -0.20 | 0.106 |
| healthcare | voting | -0.08 | 0.317 | 0.14 | 0.270 | -0.14 | 0.127 |
| healthcare | trust | 0.17 | 0.292 | -0.04 | 0.267 | **-0.29** | **0.039** |
| healthcare | fear | **-0.36** | **0.047** | -0.11 | 0.356 | 0.09 | 0.247 |
| media | community | -0.14 | 0.298 | -0.01 | 0.463 | 0.29 | 0.070 |
| media | voting | -0.16 | 0.332 | **0.37** | **0.028** | 0.06 | 0.365 |
| media | trust | 0.30 | 0.080 | 0.16 | 0.222 | -0.20 | 0.132 |
| media | fear | **-0.39** | **0.043** | 0.03 | 0.416 | 0.09 | 0.364 |
| community | voting | -0.06 | 0.297 | **-0.33** | **0.008** | -0.02 | 0.316 |
| community | trust | -0.03 | 0.431 | 0.16 | 0.348 | 0.04 | 0.640 |
| community | fear | -0.14 | 0.359 | -0.06 | 0.595 | -0.08 | 0.571 |
| voting | trust | -0.07 | 0.323 | -0.09 | 0.218 | -0.13 | 0.215 |
| voting | fear | 0.01 | 0.413 | 0.25 | 0.086 | 0.22 | 0.081 |
| trust | fear | 0.10 | 0.346 | 0.01 | 0.377 | 0.01 | 0.355 |

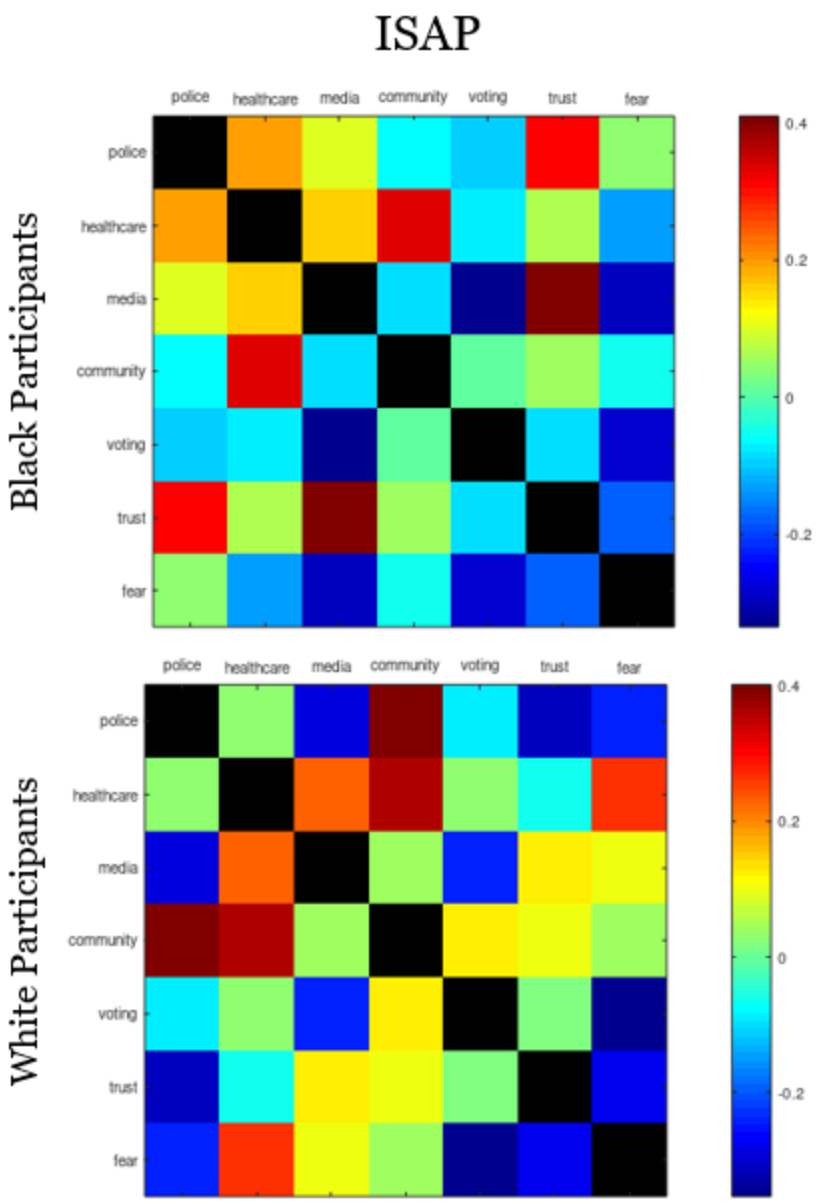


**Figure 4.3.** Heatmap of the similarity between concept-pairs across White and Black participants for the ISAP. Data were normalized within each participant and averaged across participants of the same group. The normalized data were then multiplied by -1 to invert the polarity such that positive (red) values reflect greater concept similarity, and negative (blue) values reflect concept dissimilarity.

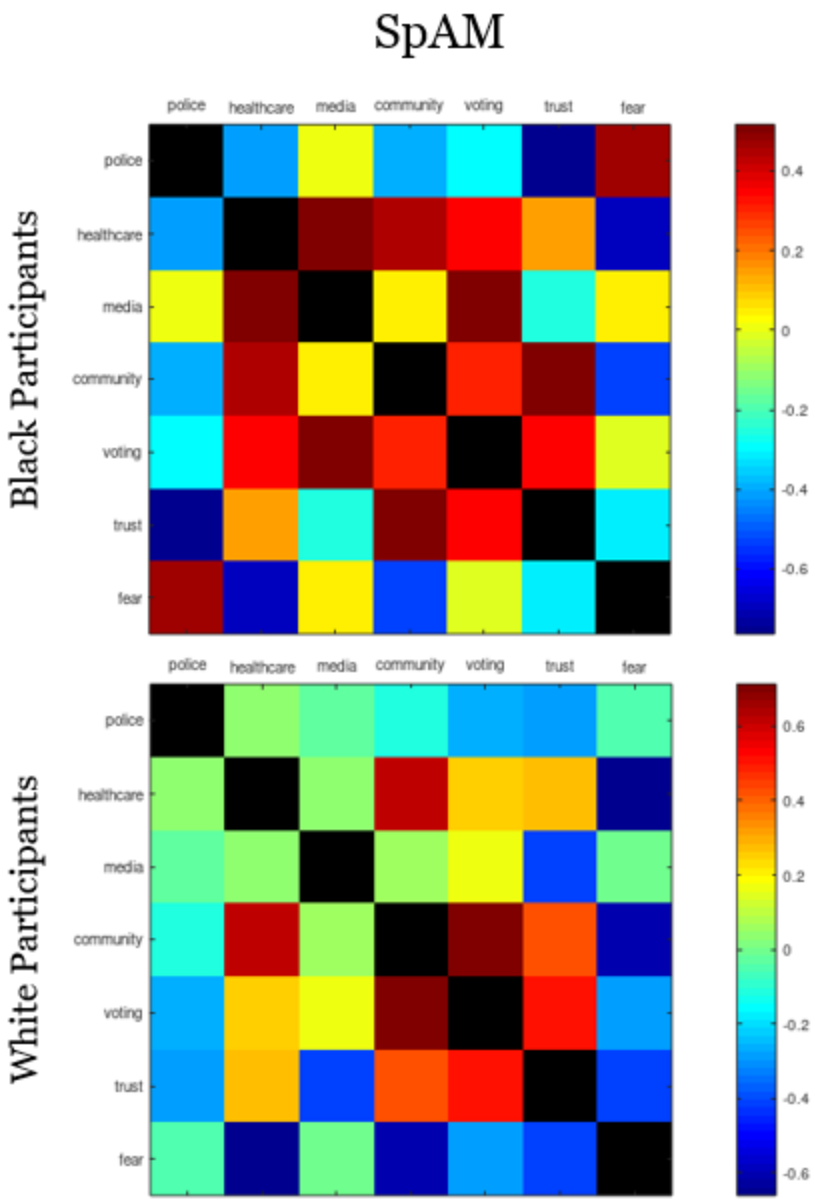


**Figure 4.4.** Heatmap of the similarity between concept-pairs across White and Black participants for the SpAM. Data were normalized within each participant and averaged across participants of the same group. The normalized data were then multiplied by -1 to invert the polarity such that positive (red) values reflect greater concept similarity, and negative (blue) values reflect concept dissimilarity.

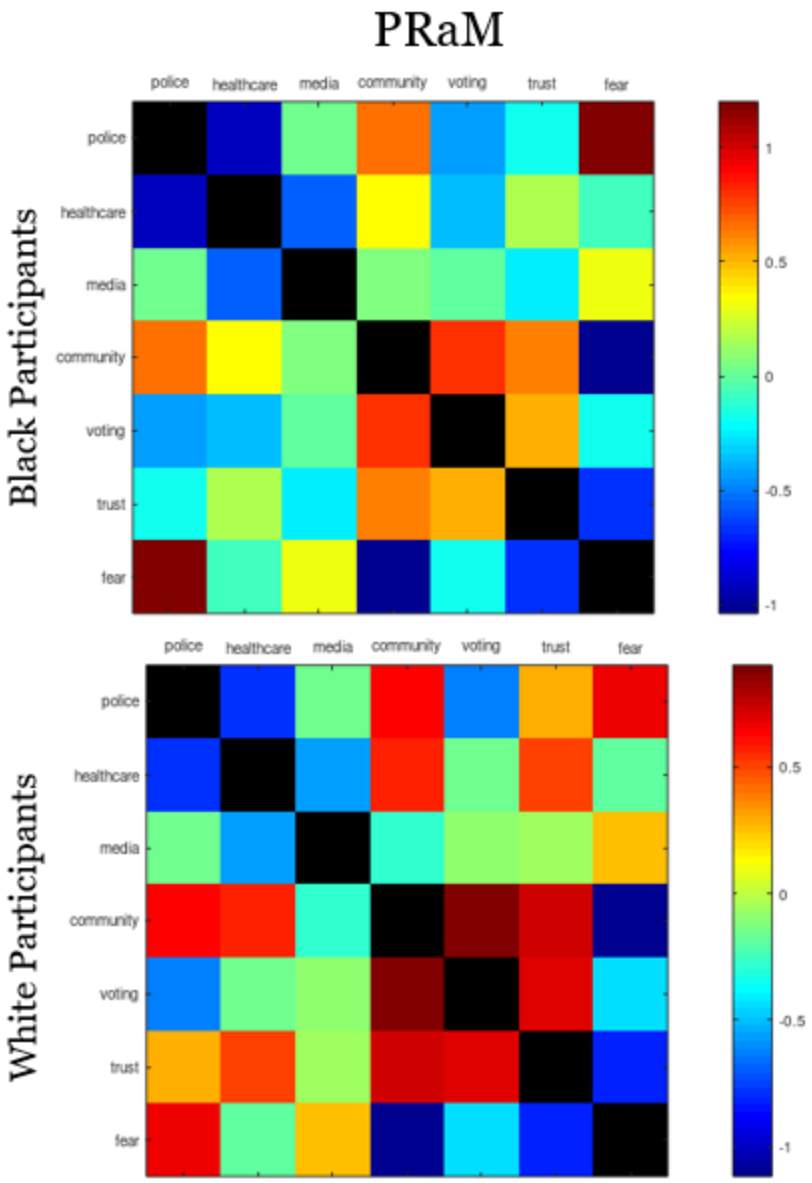


**Figure 4.5.** Heatmap of the similarity between concept-pairs across White and Black participants for the PRaM. Data were normalized within each participant and averaged across participants of the same group. The normalized data were then multiplied by -1 to invert the polarity such that positive (red) values reflect greater concept similarity, and negative (blue) values reflect concept dissimilarity.

## Discussion

The results of Study 2 reinforce finding from Study 1 suggesting the ISAP is not comparable to the SpAM and PRaM as tool for measuring concept geometry. Notable, there was a general decrease in split-half reliability across all measures compared to Study 1 (Table 4.2 and Table 4.3). This decrease suggests that the sets of concepts examined between Study 1 and Study 2 qualitatively differed. The concepts of Study 1 were predominantly concrete concepts with clear category membership and unambiguous relationships while the concepts from Study 2 were more abstract with no predetermined category structure. A further examination of the cause of this decrease in reliability will be required. Regardless, there was a significant correlation between the similarity structure of the socioenvironmental concepts across measures of concept association. Although there was a significant relationship across racial groups for the set of socioenvironmental concepts (Table 4.4), some notable differences were identified between Black and White Americans when looking at individual concept pairs.

### *Racial differences in the representation of societal environmental concepts*

There are differences in the similarity of socioenvironmental concept-pairs across Black and White racial groups. Due to the lack of reliability and validity of the ISAP the

results of the SpAM and PRaM will be the focus for interpretation (Table 4.3). Despite there being a significant correlation across the set of 21 concept pairs across Black and White participants for the SpAM and PRaM (Table 4.4), there was a double-dissociation between the association of *police* with *trust* and *police* with *fear* across Black and White racial groups for the SpAM and PRaM (Table 4.5). Black participants showed greater similarities between *police-fear* and greater dissimilarity between *police-trus*t while White participants as a group showed the inverse effects. These results are consistent with empirical studies on the differential experience of Black and White Americans' interactions with police. Black participants are more likely to experience violence by police (Plant and Peruche, 2005; Edwards et al., 2019) and more likely to have negative experiences during police traffic stops. (Voigt et al., 2017). Other significant concept-pair association include *media-voting* and *healthcare-media* for Black participants and *healthcare-trust* and *community-voting* for White participants. It is worth noting that after correcting for family-wise error the significance of the differences for each task is eliminated. Without a clear double-dissociation it is unclear what these significant relationships suggest without further analysis of the underlying representational structure.

Subsequent analyses could incorporate additional behavioral measures to predict differences in the association of socioenvironmental concepts. Moreover, the dissipation

of the significant findings after correcting for family-wise error may suggest that the current iteration of the study is underpowered. Although there is modest evidence, the elimination of significant effects after correcting for FDR suggests this study is underpowered. Moreover, despite the sample being heterogenous for variables such as age and location the samples were homogenous with respect to gender with both samples consisting of predominantly female participants. This unequal representation of gender within each group raises the question of whether these results would have differed if the samples were more equivalent with respect to gender. Although there is no significant difference in proportion life-time deaths due to police violence across White and Black women (Edwards et al., 2019), there is a significant difference in the number of experienced illegitimate police stops across White and Black women (Cochran and Warren, 2012). If the gender distributions were equivalent, then I suspect the differences in conceptual geometries between racial groups would be exacerbated. The mortality rate ratio of deaths due to police force is 2.5:1 between Black males (~100 deaths per 100,000 people) and White males (~40 deaths per 100,000 people)(Edwards et al., 2019). I suspect these dramatic differences would be reflected in the representational geometries across racial group for the *police-trust* and *police-fear* associations. Collectively these results provide an initial step towards understanding the

influence of differential experiences and treatment of systemic institutions on the conceptual association of those institutions.

# Chapter 5. General Discussion

Concepts are internal representations of characteristics of the external world with an ontological structure that associates related concepts together, like an apple tastes sweet and can sometimes be red. Concepts can instantiate our lived experiences into language and can be used to communicate our emotions and feelings towards people or institutions.

This dissertation extends upon existing research on the representation and association of concepts by examining the underlying semantic structure of abstract concept in the brain and defining a set of underlying dimensions representing the organization principles of abstract concepts, determining the commonality of those dimensions across languages, and to begin to take initial steps to measure concept representational differences across groups of individuals based on disparities of experiences of those concepts.

## *Neurosemantic dimensions of abstract meaning*

Neurally-based semantic dimensions underlying abstract concepts differ from the dimensions underlying concrete concepts. Vargas and Just (2020) investigated the fMRI activation patterns of 28 abstract concepts (e.g., *ethics*, *truth*, *spirituality*) focusing on

individual concept representation and the relationship between the activation profiles of these concept representations.

Factor analyses of the activation patterns evoked by the stimulus set revealed three underlying semantic dimensions. These dimensions corresponded to 1. the degree to which a concept was Verbally Represented; 2. whether a concept was External (or Internal) to the individual, and 3. whether the concept contained Social Content. The Verbal Representation dimension was present across all participants and was the most salient of the semantic dimensions. Concepts with large positive factor scores for this factor included compliment, faith, and ethics while concepts with large negative scores for this factor included gravity, force, and acceleration. The former three concepts seem far less perceptual than the latter three. For the Externality factor, a concept that is external is one that requires the representation of the world outside oneself and the relative non-involvement of one's own state. An internal concept is one that involves the representation of the self. At one extreme of the dimension lie concepts that are external to the self (e.g., *causality*, *sacrilege*, and *deity*). At the other extreme lie concepts that are internal to the participant (e.g., *spirituality* and *sadness*). The last semantic dimension was interpreted to correspond to Social Content. The concepts at one extreme of the dimension included pride, gossip, and equality while the concepts at the

other extreme included heat, necessity, and multiplication. Together these semantic dimensions underlie the neural representations of these 28 abstract concepts.

One surprising finding was that the regions associated with the Verbal Representation dimension were the same as those found in the meta-analysis conducted by Wang and colleagues (2010) that contrasted the activation between concrete and abstract concepts. Activation in LIFG (a region clearly involved in verbal processing) was evoked by concepts such as faith and truth while the left supramarginal gyrus (LSMG) and left lateral occipital complex (LOC), both of which are involved in different aspects of visuospatial processing, were associated with concepts such as gravity and heat.

Moreover, the output of the factor analysis (i.e., factor scores) for the Verbal Representation factor also suggested that the abstractness of the neural patterning in these regions for an individual concept is represented as a point on a continuum between language systems and perceptual processing systems. This interpretation corresponds to the intuition that abstractness is not a binary construct but rather a gradient-like translation of a concept into a more verbal encoding. This conclusion is somewhat surprising given that the set of 28 concepts are all qualitatively abstract, in that they have no direct perceptual referent. The amount of activation in LIFG evoked by a given abstract concept corresponds to its Verbal Representation factor score.

*Commonality of individual concrete and abstract concepts across people*

Several discussions above have noted that the semantic factor structure was similar across participants, but that understates one of the most interesting findings in the area of neural representations of concepts. The surprising finding is that the neural representations of all of the concepts studied so far are rather similar across people. This section below focuses on the commonality of individual concept representations across individuals. The general approach to quantitatively evaluating the commonality of individual concept representations is to train a machine learning classifier on the labeled activation data of all but one participant for a given set of concepts, and then to classify or make predictions concerning the concept representations of the left-out individual. In a cross-validation protocol, this process is repeated with a different person left out on each iteration. The accuracies of the predictions are then averaged across iterations. This averaged accuracy measures the commonality of a set of concept representations.

This approach has shown that there is considerable commonality of the neural representations of concepts across healthy participants. The commonality was present for concrete, abstract, and hybrid concepts. Decoding accuracies across participants

were high and approximately equivalent for concrete, abstract, and hybrid concepts (i.e., mean rank accuracy = 0.72 for concrete concepts (Just et al., 2010); 0.74 for abstract concepts (Vargas and Just, 2020); 0.71 for physics concepts (Mason and Just, 2016); 0.7 for emotion concepts (Kassam et al., 2013); and 0.77 for social concepts (Just et al., 2014).

Although a large proportion of the concepts in a brain reading study are accurately predicted across participants, there are always a few items at the negative tail of the accuracy distribution, and it would be interesting to know if the items with lower across-participant commonalities had some distinguishing properties. In a study of sentence decoding across three languages (Portuguese, Mandarin, and English), Yang et al. (2017) found lower across-language, across participant decoding accuracies for concepts that are more abstract and related to social and mental activities (e.g., happy, negotiation, artist). They attributed this lower degree of commonality across languages of such items to some abstract and socially related concept domains being more culturally-determined.

For the set of 28 abstract concepts presented in the Vargas and Just (2020) study, the concepts which were more prototypically abstract (e.g., sacrilege and contract) were somewhat less accurately predicted across participants than concepts that tend to be more hybrid (e.g., *force* and *pride*). However, there were a number of

exceptions to this trend. For example, concepts such as anger and gossip were less well predicted than others across participants (although still with an above-chance accuracy), and these concepts tended to be highly instantiable. By contrast, concepts such as necessity and causality, which are highly verbally represented, were more accurately predicted across participants.

When examining the neural representation of the same 28 abstract concepts across languages and cultures by looking at the fMRI-measured activation of native English and native Mandarin speakers, factor analyses revealed a set of common neurosemantic dimensions that constitute the basis for the representation of abstract concepts across languages: verbal representation, internality–externality, social content, and rule-based content (Vargas and Just, 2022). The successful cross-language classification suggests that the underlying semantic dimensions provide a sufficient basis for decoding abstract word concepts across languages. Although the neural dimensions used for representing abstract concepts are common across languages, differences in the meaning of some individual concepts can be accommodated in terms of differential salience of particular dimensions. These semantic dimensions constitute a set of neurocognitive resources for abstract concept representation within a larger set of regions responsible for general semantic processing. These results raise an interesting theoretical and psychological question regarding the role of neural language systems,

particularly LIFG, in the verbal representation of abstract concepts. That is, what does it mean, neurally and psychologically, for an abstract concept to be verbally represented?

*Abstract concepts as verbal representations*

What does it mean for an abstract concept to be represented in regions involved in verbal processing and to evoke activation in LIFG? When LIFG is artificially lesioned through the repeated use of transcranial magnetic stimulation (TMS), otherwise healthy participants show a 150 ms slower response time for comprehending abstract concepts (e.g., *chance*) (Hoffman, et al., 2010). This same TMS-based lesioning procedure showed no influence in the amount of time needed to respond to concrete concepts (e.g., *apple*). However, these differences in the impact of TMS were nullified when the abstract concepts were presented within a context (e.g., "*You don't stand a chance*"). These results suggest that the abstractness of a concept is dependent on whether it requires integration of meaning across multiple contexts (Crutch and Warrington 2005; Crutch and Warrington, 2010; Hoffman 2016; Hayes and Kraemer, 2017). Moreover, the LIFG seems to be involved in the context-dependent integration of meaning.

Given that LIFG appears to be involved in the contextualization of the meaning of abstract concepts (Hoffman et al., 2010) and that the magnitude of activation in LIFG is

directly proportional to the degree that it is verbally represented (Vargas and Just, 2020), taken together these results suggest that the activation in LIFG reflects the magnitude of mental activity required to contextualize the meaning of a lexical concept. LIFG has been shown to elicit greater activation for sentence-level representations as compared to word-level concepts (Xu et al., 2005). It may be the case that the central cognitive mechanism underlying the neural activation in LIFG represents the integration of meaning across multiple representations in order to form a new representation that is a product of its components. That is, the components of meaning of *apple* require less computation (in LIFG) to generate a composite representation than the concept of *chance*. Also, providing a context for *chance,* as in "*You don't stand a chance",* reduced the cognitive workload by providing a more explicit version of its meaning. A similar mechanism can account for the greater activation in LIFG for sentences than for individual words, because constructing a sentence-level representation requires combining the meanings of individual concept representations in a mutually context-constraining way.

As previously discussed, another region involved in the integrating of meaning for concepts is the anterior temporal lobe (ATL). ATL has been implicated in the integration of semantic features to form a composite representation of object concepts (Coutanche and Thompson-Schill, 2015). However, unlike LIFG, ATL does not appear to

differentiate between abstract concepts that vary based on the degree that they are verbally represented (as defined by their factor scores in Vargas and Just (2020)).

In sum, the integration of abstract concept representations with other concepts in a sentence seems to require additional computation. However, it is unclear whether these integrating computations are processing some episodic contexts (as suggested by the results of Hoffman et al., 2010), or some specific concept representations, or use some more general amodal representational format.

*Incompatibility between behavioral and neural conceptual geometries*

It is unclear whether it would be possible to directly map the semantic organization of concepts across behavioral and neural environments. To our knowledge, there has not been an attempt to replicate behavioral conceptual geometries from neural activation. The mapping between behavioral and neural representational spaces is complicated by the exhaustive nature of neural data. The neural evocation of a concept constitutes all aspects of its representation – from activation related to the early visual processing of viewing light corresponding to the stimulus word on a computer screen to activation associated with executive function to maintain attention to task demands. This is further complicated by the distributed and piecemeal nature of neural activation

patterns (Meyer and Damasio, 2009). As shown in the two projects examining the neural representation of abstract concepts, each semantic organization principle is represented not as a single region but rather an interplay between a distributed network of associated brain regions. The advantage of examining the neural organization principles of concepts lies in the method's ability to identify the mechanistic processes that contribute to the evocation of a set of concepts. Behaviorally measured conceptual geometries are shielded from these complexities but are susceptible to other possible confounds such as response bias to experimental tasks. Where neural concept representations reflect the composition of concept, I suspect behavioral geometries reflect inter-concept relations of the composed concept. Thus, I suspect behavioral and neural geometries reflect distinct constructs. Behavioral measures of concept geometry capture how people think about a concept while the neural geometries reflect what goes into thinking about a concept.

### *Preliminary evidence for differences in concept representation based on environmental factors*

Although there has been a tremendous amount of research examining the influence of environmental factors on biased attributions towards racial group there has been less research examining whether concept associations reflect disparities in the

experience of those concepts. The bias-of-crowds model suggests that implicit concept associations reflect environmental contextual and situation factors including situational and chronic experiences with those concepts. The vast and well documented empirical research on racial disparities between Black and White Americans provides this chronic experiential context for the socioenvironmental concepts where these disparities persist. The wisdom-of-crowds phenomena suggests the transience of concept representations dissipates when examining over a large group of people. This demand for larger sample sizes is prohibitive for the examination of differences in representation of socioenvironmental concepts using an fMRI methodology. Thus, to gain an initial understanding of the influence of concept differences, a large-scale online behavioral study was conducted. By examining behaviorally-measured associations between socioenvironmental concepts pertaining to institutions where these disparities are present, preliminary evidence suggesting a double-dissociation in the represent of the concept *police*. Black participants seem to show less trust and greater fear towards the police relative to White participants. Although this finding is modest in isolation, follow up research and analyses could provide insight into the geometry of socioenvironmental concepts now that the behavioral tasks for measuring concept associations has been established.

Generally speaking, differences in treatment on a systemic level across Black and White Americans are likely driving the differences in conceptual geometries. This systemic influence creates a persistent contextual mechanism for differential experience of concepts central to living in American society. This widespread mechanism makes it difficult to pinpoint which exact facets or contextual factors are driving the differences. This raises the question of how differences in conceptual geometries could be better measured to capture differences reflected in the racial disparities literature? With a limited number of attributions in Study 2 it is unclear what attributional mechanisms could be underlying concepts where racial disparities exist. It is likely the case that neither *fear* nor *trust* are appropriate for capturing how people feel towards certain concepts (e.g., *community*). Moreover, having additional attribution could provide possible explanations for significant non-attributional associations between concept pairs (e.g., *police-healthcare* and *community-voting* are significantly more associated for White participants than Black participants)(Table 4.5). Insight into which other attributes could be included into the concept set could be gained from the study of emotions. Theoretical work on basic emotions suggests that 6 basic emotions underlie human cognition (i.e., joy, love, sadness, fear, anger, and surprise)(Shaver et al., 1987). The similarity ratings of each of these basic emotions with socioenvironmental concepts

could provide insight into the nuanced and complex emotional association people have towards these concepts.

*Future direction*

Taken collectively, this dissertation provided a conceptual springboard for the investigation of social factors on concept associations related to societal institutions. The first two projects of this dissertation provide a neurosemantic basis for the underlying semantic structure of abstract concepts. With this basis for the generalizability of abstract concepts across individuals, we can begin to ask questions about abstract concepts in the context of our social world. However, in the spirit of the wisdom-of-crowds phenomena special consideration needs to be taken to not draw inferences about a social identity (such a race) based on a small sample of participants. Future behavioral work would need to collect large samples using psychometrically sound measurements and techniques (such as the SpAM and PRaM). Using population-level data is the key to understanding the semantic structure underling these concept associations. In addition to population-level data, larger concept set sizes are needed to extrapolate more nuanced dynamics constituting societal concept geometries. Moreover, further work needs incorporate measurements of possible contributing mechanisms

such as social dominance orientation and the importance of someone's identity to their association.

Further work could also investigation the relationship between societal concept geometries and behavior, specifically, language use and decision making. The relationship between concept geometries and language could be linked by examining social media data through a natural language inference (NLI) perspective. This would potentially illuminate discrepancies in how people view their societal world and how they choose to communicate in a way that reflects associations between societal institutions. Moreover, by investigating the link between concept associations and decision-making, specific links can be made between thought and people's differential decisions to reward or punish in-group versus outgroup members. The goal is to strive for an understanding of the composition and contributions to an individual's world view as they navigate a social world.

## References


Anderson, A. J., Kiela, D., Clark, S., & Poesio, M. (2017). Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, *5*, 17–30. https://direct.mit.edu/tacl/article-abstract/doi/10.1162/tacl_a_00043/43388

Anderson, A. J., Murphy, B., & Poesio, M. (2014). Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. *Journal of Cognitive Neuroscience*, *26*(3), 658–681. https://doi.org/10.1162/jocn_a_00508

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Ash, M., & Boyce, J. K. (2018). Racial disparities in pollution exposure and employment at US industrial facilities. *Proceedings of the National Academy of Sciences*, *115*(42), 10636–10641. https://doi.org/10.1073/pnas.1721640115

Badre, D., Poldrack, R. A., Paré-Blagoev, E. J., Insler, R. Z., & Wagner, A. D. (2005). Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. *Neuron*, *47*(6), 907–918. https://doi.org/10.1016/j.neuron.2005.07.023

Barsalou, L. W. (1999). Perceptual symbol systems. *The Behavioral and Brain Sciences*, *22*(4), 577–609; discussion 610–660. https://doi.org/10.1017/s0140525x99002149

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *358*(1435), 1177–1187. https://doi.org/10.1098/rstb.2003.1319

Bauer, A. J., & Just, M. A. (2017). A brain-based account of "basic-level" concepts. *NeuroImage*, *161*, 196–205. https://doi.org/10.1016/j.neuroimage.2017.08.049

Benn, Y., Zheng, Y., Wilkinson, I. D., Siegal, M., & Varley, R. (2012). Language in calculation: a core mechanism? *Neuropsychologia*, *50*(1), 1–10. https://doi.org/10.1016/j.neuropsychologia.2011.09.045

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 , 19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Rao, S. M., & Cox, R. W. (1999). Conceptual processing during the conscious resting state. A functional MRI study. *Journal of Cognitive Neuroscience*, *11*(1), 80–95. https://doi.org/10.1162/089892999563265

Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, *17*(6), 905–917. https://doi.org/10.1162/0898929054021102

Bird, C. M., Keidel, J. L., Ing, L. P., Horner, A. J., & Burgess, N. (2015). Consolidation of Complex Events via Reinstatement in Posterior Cingulate Cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(43), 14426–14434. https://doi.org/10.1523/JNEUROSCI.1774-15.2015

Brown-Iannuzzi, J. L., Dotsch, R., Cooley, E., & Payne, B. K. (2017). The Relationship Between Mental Representations of Welfare Recipients and Attitudes Toward Welfare. *Psychological Science*, *28*(1), 92–103. https://doi.org/10.1177/0956797616674999

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Buchweitz, A., Shinkareva, S. V., Mason, R. A., Mitchell, T. M., & Just, M. A. (2012). Identifying bilingual semantic neural representations across languages. *Brain and Language*, *120*(3), 282–289. https://doi.org/10.1016/j.bandl.2011.09.003

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS One, 5*(6), e10729. https://doi.org/10.1371/journal.pone.0010729

Chen, J. L., Zatorre, R. J., & Penhune, V. B. (2006). Interactions between auditory and dorsal premotor cortex during synchronization to musical rhythms. *NeuroImage, 32*(4), 1771–1781. https://doi.org/10.1016/j.neuroimage.2006.04.207

Clarke, A., Taylor, K. I., Devereux, B., Randall, B., & Tyler, L. K. (2013). From perception to conception: how meaningful objects are processed over time. *Cerebral Cortex , 23*(1), 187–197. https://doi.org/10.1093/cercor/bhs002

Cochran, J. C., & Warren, P. Y. (2012). Racial, ethnic, and gender differences in perceptions of the police: The salience of officer race within the context of racial profiling. Journal of contemporary criminal justice, 28(2), 206-227.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Collins, D. L., Neelin, P., Peters, T. M., & Evans, A. C. (1994). Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of Computer Assisted Tomography, 18*(2), 192–205. https://www.ncbi.nlm.nih.gov/pubmed/8126267

Coutanche, M. N., & Thompson-Schill, S. L. (2015). Creating Concepts from Converging Features in Human Cortex. *Cerebral Cortex , 25*(9), 2584–2593. https://doi.org/10.1093/cercor/bhu057

Crutch, S. J., & Warrington, E. K. (2005). Abstract and concrete concepts have structurally different representational frameworks. *Brain: A Journal of Neurology*, *128*(Pt 3), 615–627. https://doi.org/10.1093/brain/awh349

Crutch, S. J., & Warrington, E. K. (2010). The differential dependence of abstract and concrete words upon associative and similarity-based information: Complementary semantic interference and facilitation effects. *Cognitive Neuropsychology*, *27*(1), 46–71. https://doi.org/10.1080/02643294.2010.491359

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science. American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9

Dehghani, M., Boghrati, R., Man, K., Hoover, J., Gimbel, S. I., Vaswani, A., Zevin, J. D., Immordino-Yang, M. H., Gordon, A. S., Damasio, A., & Kaplan, J. T. (2017). Decoding the neural representation of story meanings across languages. *Human Brain Mapping*, *38*(12), 6096–6106. https://doi.org/10.1002/hbm.23814

Dreyer, F. R., & Pulvermüller, F. (2018). Abstract semantics in the motor system?--An event-related fMRI study on passive reading of semantic word categories carrying abstract emotional and mental meaning. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *100*, 52–70. https://doi.org/10.1016/j.cortex.2017.10.021

Edwards, F., Lee, H., & Esposito, M. (2019). Risk of being killed by police use of force in the United States by age, race–ethnicity, and sex. *Proceedings of the National Academy of Sciences*, *116*(34), 16793–16798. https://doi.org/10.1073/pnas.182120411

Firebaugh, G., & Acciai, F. (2016). For blacks in America, the gap in neighborhood poverty has declined faster than segregation. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(47), 13372–13377. https://doi.org/10.1073/pnas.1607220113

Fugelsang, J. A., & Dunbar, K. N. (2005). Brain-based mechanisms underlying complex causal thinking. *Neuropsychologia*, *43*(8), 1204–1213. https://doi.org/10.1016/j.neuropsychologia.2004.10.012

Fugelsang, J. A., Roser, M. E., Corballis, P. M., Gazzaniga, M. S., & Dunbar, K. N. (2005). Brain mechanisms underlying perceptual causality. *Brain Research. Cognitive Brain Research*, *24*(1), 41–47. https://doi.org/10.1016/j.cogbrainres.2004.12.001

Fuhrman, O., McCormick, K., Chen, E., Jiang, H., Shu, D., Mao, S., & Boroditsky, L. (2011). How linguistic and cultural forces shape conceptions of time: English and Mandarin time in 3D. *Cognitive Science*, *35*(7), 1305–1328. https://doi.org/10.1111/j.1551-6709.2011.01193.x

Gay, C. (2004). Putting Race in Context: Identifying the Environmental Determinants of Black Racial Attitudes. *The American Political Science Review*, *98*(4), 547–562. https://doi.org/10.1017/S0003055404041346

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *26*(4), 381–386. https://doi.org/10.3758/BF03204653

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., Hahn, A., Hehman, E., Hofmann, W., Hughes, S., Hussey, I., Jordan, C., Kirby, T. A., Lai, C. K., Lang, J. W. B., Lindgren, K. P., Maison, D., Ostafin, B. D., Rae, J. R., … Wiers, R. W. (2022). Best research practices for using the Implicit Association Test. *Behavior Research Methods*, *54*(3), 1161–1180. https://doi.org/10.3758/s13428-021-01624-3

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social*

*Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037//0022-3514.74.6.1464

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10-11), 1409–1422. https://doi.org/10.1016/s0042-6989(01)00073-6

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, *27*, 649–677. https://doi.org/10.1146/annurev.neuro.27.070203.144220

Guterstam, A., Björnsdotter, M., Gentile, G., & Ehrsson, H. H. (2015). Posterior cingulate cortex integrates the senses of self-location and body ownership. *Current Biology: CB*, *25*(11), 1416–1425. https://doi.org/10.1016/j.cub.2015.03.059

Han, S., & Northoff, G. (2008). Culture-sensitive neural substrates of human cognition: a transcultural neuroimaging approach. *Nature Reviews. Neuroscience*, *9*(8), 646–654. https://doi.org/10.1038/nrn2456

Harpaintner, M., Sim, E.-J., Trumpp, N. M., Ulrich, M., & Kiefer, M. (2020). The grounding of abstract concepts in the motor and visual system: An fMRI study. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *124*, 1–22. https://doi.org/10.1016/j.cortex.2019.10.014

Hauk, O., & Pulvermüller, F. (2004). Neurophysiological distinction of action words in the fronto-central cortex. *Human Brain Mapping*, *21*(3), 191–201. https://doi.org/10.1002/hbm.10157

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Hayes, J. C., & Kraemer, D. J. M. (2017). Grounded understanding of abstract concepts: The case of STEM learning. *Cognitive Research: Principles and Implications*, *2*(1), 7. https://doi.org/10.1186/s41235-016-0046-z

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews. Neuroscience*, *7*(7), 523–534. https://doi.org/10.1038/nrn1931

Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., Foels, R., & Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO$_7$ scale. *Journal of Personality and Social Psychology*, *109*(6), 1003–1028. https://doi.org/10.1037/pspi0000033

Hoffman, P. (2016). The meaning of "life" and other abstract words: Insights from neuropsychology. In *Journal of Neuropsychology* (Vol. 10, Issue 2, pp. 317–343).

https://doi.org/10.1111/jnp.12065

Hoffman, P., Jefferies, E., & Lambon Ralph, M. A. (2010). Ventrolateral prefrontal cortex plays an executive regulation role in comprehension of abstract words: convergent neuropsychological and repetitive TMS evidence. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(46), 15450–15456. https://doi.org/10.1523/JNEUROSCI.3783-10.2010

Hofmann, W., & Baumert, A. (2010). Immediate affect as a basis for intuitive moral judgement: An adaptation of the affect misattribution procedure. *Cognition and Emotion*, *24*(3), 522–535. https://doi.org/10.1080/02699930902847193

Hofmann, W., van Koningsbruggen, G. M., Stroebe, W., Ramanathan, S., & Aarts, H. (2010). As pleasure unfolds. Hedonic responses to tempting food. *Psychological Science*, *21*(12), 1863–1870. https://doi.org/10.1177/0956797610389186

Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of SpAM: a fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology. General*, *142*(1), 256–281. https://doi.org/10.1037/a0028860

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

453–458. https://doi.org/10.1038/nature17637

Jefferies, E., Patterson, K., Jones, R. W., & Lambon Ralph, M. A. (2009). Comprehension of concrete and abstract words in semantic dementia. In *Neuropsychology* (Vol. 23, Issue 4, pp. 492–499). https://doi.org/10.1037/a0015452

Just, M. A., Cherkassky, V. L., Aryal, S., & Mitchell, T. M. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS One*, *5*(1), e8622. https://doi.org/10.1371/journal.pone.0008622

Just, M. A., Cherkassky, V. L., Buchweitz, A., Keller, T. A., & Mitchell, T. M. (2014). Identifying autism from neural representations of social interactions: neurocognitive markers of autism. *PloS One*, *9*(12), e113879. https://doi.org/10.1371/journal.pone.0113879

Just, M. A., Pan, L., Cherkassky, V. L., McMakin, D. L., Cha, C., Nock, M. K., & Brent, D. (2017). Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature Human Behaviour*, *1*, 911–919. https://doi.org/10.1038/s41562-017-0234-y

Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G., & Just, M. A. (2013). Identifying Emotions on the Basis of Neural Activation. *PloS One*, *8*(6), e66032.

https://doi.org/10.1371/journal.pone.0066032

Kendi, I. X. (2016). *Stamped from the Beginning: The Definitive History of Racist Ideas in America*. PublicAffairs.

Kim, S. J., & Bostwick, W. (2020). Social Vulnerability and Racial Inequality in COVID-19 Deaths in Chicago. *Health Education & Behavior: The Official Publication of the Society for Public Health Education*, *47*(4), 509–513. https://doi.org/10.1177/1090198120929677

Klasen, M., Kenworthy, C. A., Mathiak, K. A., Kircher, T. T. J., & Mathiak, K. (2011). Supramodal representation of emotions. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(38), 13635–13643. https://doi.org/10.1523/JNEUROSCI.2833-11.2011

Kober, H., Barrett, L. F., Joseph, J., Bliss-Moreau, E., Lindquist, K., & Wager, T. D. (2008). Functional grouping and cortical–subcortical interactions in emotion: A meta-analysis of neuroimaging studies. *NeuroImage*, *42*(2), 998–1031. https://doi.org/10.1016/j.neuroimage.2008.03.059

Koch, A., Speckmann, F., & Unkelbach, C. (2020). Q-SpAM: How to Efficiently Measure Similarity in Online Research. *Sociological Methods & Research*, 0049124120914937.

https://doi.org/10.1177/0049124120914937

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. *Journal of Experimental Psychology. General*, *140*(1), 14–34. https://doi.org/10.1037/a0021446

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Lai, V. T., & Boroditsky, L. (2013). The immediate and chronic influence of spatio-temporal metaphors on the mental representations of time in english, mandarin, and mandarin-english speakers. *Frontiers in Psychology*, *4*, 142.

https://doi.org/10.3389/fpsyg.2013.00142

Lambon Ralph, M. A. (2014). Neurocognitive insights on conceptual knowledge and its breakdown. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1634), 20120392. https://doi.org/10.1098/rstb.2012.0392

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295x.104.2.211

Lazri, A. M., & Konisky, D. M. (2019). Environmental attitudes across race and ethnicity. *Social Science Quarterly*, *100*(4), 1039–1055. https://doi.org/10.1111/ssqu.12626

Lee, K. M., Lindquist, K. A., & Payne, B. K. (2018). Constructing bias: Conceptualization breaks the link between implicit bias and fear of Black Americans. *Emotion*, *18*(6), 855–871. https://doi.org/10.1037/emo0000347

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Liuzzi, A. G., Aglinskas, A., & Fairhall, S. L. (2020). General and feature-based semantic representations in the semantic network. *Scientific Reports*, *10*(1), 8931.

https://doi.org/10.1038/s41598-020-65906-0

Lord, F. M., & Novick, M. R. (2008). *Statistical Theories of Mental Test Scores*. IAP.

Maddock, R. J., Garrett, A. S., & Buonocore, M. H. (2003). Posterior cingulate cortex activation by emotional words: fMRI evidence from a valence decision task. *Human Brain Mapping*, *18*(1), 30–41. https://doi.org/10.1002/hbm.10075

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*, 25–45. https://doi.org/10.1146/annurev.psych.57.102904.190143

Mason, R. A., & Just, M. A. (2016). Neural Representations of Physics Concepts. *Psychological Science*, *27*(6), 904–913. https://doi.org/10.1177/0956797616641941

Mason, R. A., & Just, M. A. (2020). Neural Representations of Procedural Knowledge. *Psychological Science*, *31*(6), 729–740. https://doi.org/10.1177/0956797620916806

Meyer, K., & Damasio, A. (2009). Convergence and divergence in a neural architecture for recognition and memory. *Trends in Neurosciences*, *32*(7), 376–382. https://doi.org/10.1016/j.tins.2009.04.002

Millett, G. A., Jones, A. T., Benkeser, D., Baral, S., Mercer, L., Beyrer, C., Honermann, B., Lankiewicz, E., Mena, L., Crowley, J. S., Sherwood, J., & Sullivan, P. S. (2020).

Assessing differential impacts of COVID-19 on black communities. *Annals of Epidemiology*, *47*, 37–44. https://doi.org/10.1016/j.annepidem.2020.05.003

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

Mozaffarian, D., Benjamin, E. J., Go, A. S., & Arnett, D. K. (2015). Heart disease and stroke statistics—2015 update: a report from the American Heart Association. *Circulation*. https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000000152

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience*, *4*(1), 101–109. https://doi.org/10.1093/scan/nsn044

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

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology: CB*, *21*(19), 1641–1646. https://doi.org/10.1016/j.cub.2011.08.031

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice. *Psychological Inquiry, 28*(4), 233–248. https://doi.org/10.1080/1047840X.2017.1335568

Payne, K., & Lundberg, K. (2014). The affect misattribution procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass, 8*(12), 672–686. https://doi.org/10.1111/spc3.12148

Pecher, D., Boot, I., & Van Dantzig, S. (2011). Abstract concepts: Sensory-motor grounding, metaphors, and beyond. In *Psychology of learning and motivation* (Vol. 54, pp. 217–248). Elsevier. https://doi.org/10.1016/B978-0-12-385527-5.00007-3

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://aclanthology.org/D14-1162.pdf

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

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage,*

*16*(2), 331–348. https://doi.org/10.1006/nimg.2002.1087

Plant, E. A., & Peruche, B. M. (2005). The consequences of race for police officers' responses to criminal suspects. *Psychological Science*, *16*(3), 180–183. https://doi.org/10.1111/j.0956-7976.2005.00800.x

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews. Neuroscience*, *18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing auto- matic and controlled components of attitudes from direct and indirect measurement methods. Journal of Experimental Social Psychology, 44 (2), 386–396. doi:10.1016/j.jesp.2006.12.008

Schuck, A. M., & Rosenbaum, D. P. (2005). Global and neighborhood attitudes toward the police: Differentiation by race, ethnicity and type of contact. *Journal of Quantitative Criminology*, *21*(4), 391–418. https://doi.org/10.1007/s10940-005-7356-5

Sellers, R. M., Rowley, S. A. J., Chavous, T. M., Shelton, J. N., & Smith, M. A. (1997). Multidimensional Inventory of Black Identity: A preliminary investigation of reliability and constuct validity. *Journal of Personality and Social Psychology*, *73*(4), 805–815. https://doi.org/10.1037/0022-3514.73.4.805

Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. Journal of personality and social psychology, 52(6), 1061.

Silani, G., Lamm, C., Ruff, C. C., & Singer, T. (2013). Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *33*(39), 15466–15476. https://doi.org/10.1523/JNEUROSCI.1488-13.2013

Smith, T. W., Davern, M., Freese, J., & Hout, M. (2019). General social survey, 2018 [Data file]. Chicago, IL: NORC at the University of Chicago. Retrieved from http://gssdataexplorer.norc.org.

Surowiecki, J. (2004). *The Wisdom Of Crowds*. Anchor Books.

Tai, D. B. G., Shah, A., Doubeni, C. A., Sia, I. G., & Wieland, M. L. (2021). The Disproportionate Impact of COVID-19 on Racial and Ethnic Minorities in the United States. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *72*(4), 703–706. https://doi.org/10.1093/cid/ciaa815

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & 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), 273–289. https://doi.org/10.1006/nimg.2001.0978

Umberson, D., Olson, J. S., Crosnoe, R., Liu, H., Pudrovska, T., & Donnelly, R. (2017). Death of family members as an overlooked source of racial disadvantage in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(5), 915–920. https://doi.org/10.1073/pnas.1605599114

van Ryn, M., Burgess, D. J., Dovidio, J. F., Phelan, S. M., Saha, S., Malat, J., Griffin, J. M., Fu, S. S., & Perry, S. (2011). The impact of racism on clinician cognition, behavior, and clinical decision making. *Du Bois Review: Social Science Research on Race*, *8*(1), 199–218. https://doi.org/10.1017/S1742058X11000191

Vargas, R., & Just, M. A. (2020). Neural Representations of Abstract Concepts: Identifying Underlying Neurosemantic Dimensions. *Cerebral Cortex* , *30*(4), 2157–2166. https://doi.org/10.1093/cercor/bhz229

Vargas, R. & Just, M. (2021). The Neural Representation of Concrete and Abstract Concepts. In Barbey, A.K., Karama, S. & Haier, R.J. (Eds.) The Cambridge Handbook of Intelligence and Cognitive Neuroscience.

Vargas, R., & Just, M. A. (2022). Similarities and differences in the neural representations of abstract concepts across English and Mandarin. *Human Brain Mapping*.

https://doi.org/10.1002/hbm.25844

Verheyen, S., Voorspoels, W., Vanpaemel, W., & Storms, G. (2016). Caveats for the spatial arrangement method: Comment on Hout, Goldinger, and Ferguson (2013) [Review of *Caveats for the spatial arrangement method: Comment on Hout, Goldinger, and Ferguson (2013)*]. *Journal of Experimental Psychology. General*, *145*(3), 376–382. https://doi.org/10.1037/a0039758

Verheyen, S., White, A., & Storms, G. (2020). A Comparison of the Spatial Arrangement Method and the Total-Set Pairwise Rating Method for Obtaining Similarity Data in the Conceptual Domain. *Multivariate Behavioral Research*, 1–28. https://doi.org/10.1080/00273171.2020.1857216

Vigliocco, G., Kousta, S.-T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: the role of emotion. *Cerebral Cortex* , *24*(7), 1767–1777. https://doi.org/10.1093/cercor/bht025

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., Jurgens, D., Jurafsky, D., & Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(25), 6521–6526. https://doi.org/10.1073/pnas.1702413114

Vuletich, H. A., & Payne, B. K. (2019). Stability and Change in Implicit Bias. *Psychological Science*, *30*(6), 854–862. https://doi.org/10.1177/0956797619844270

Wang, J., Baucom, L. B., & Shinkareva, S. V. (2013). Decoding abstract and concrete concept representations based on single-trial fMRI data. *Human Brain Mapping*, *34*(5), 1133–1147. https://doi.org/10.1002/hbm.21498

Wang, J., Conder, J. A., Blitzer, D. N., & Shinkareva, S. V. (2010). Neural representation of abstract and concrete concepts: a meta-analysis of neuroimaging studies. *Human Brain Mapping*, *31*(10), 1459–1468. https://doi.org/10.1002/hbm.20950

Wang, X., Wu, W., Ling, Z., Xu, Y., Fang, Y., Wang, X., Binder, J. R., Men, W., Gao, J.-H., & Bi, Y. (2018). Organizational Principles of Abstract Words in the Human Brain. *Cerebral Cortex* , *28*(12), 4305–4318. https://doi.org/10.1093/cercor/bhx283

Xu, J., Kemeny, S., Park, G., Frattali, C., & Braun, A. (2005). Language in context: emergent features of word, sentence, and narrative comprehension. *NeuroImage*, *25*(3), 1002–1015. https://doi.org/10.1016/j.neuroimage.2004.12.013

Xu, X., & Li, J. (2020). Concreteness/abstractness ratings for two-character Chinese words in MELD-SCH. *PloS One*, *15*(6), e0232133. https://doi.org/10.1371/journal.pone.0232133

Yang, Y., Wang, J., Bailer, C., Cherkassky, V., & Just, M. A. (2017a). Commonalities and differences in the neural representations of English, Portuguese, and Mandarin sentences: When knowledge of the brain-language mappings for two languages is better than one. *Brain and Language*, *175*, 77–85. https://doi.org/10.1016/j.bandl.2017.09.007

Yang, Y., Wang, J., Bailer, C., Cherkassky, V., & Just, M. A. (2017b). Commonality of neural representations of sentences across languages: Predicting brain activation during Portuguese sentence comprehension using an English-based model of brain function. *NeuroImage*, *146*, 658–666. https://doi.org/10.1016/j.neuroimage.2016.10.029

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635