Cerebral Cortex, 2019;9: 1–10

doi: 10.1093/cercor/bhz229
Advance Access Publication Date:
Original Article

ORIGINAL ARTICLE

Neural Representations of Abstract Concepts: Identifying Underlying Neurosemantic Dimensions

Robert Vargas and Marcel Adam Just*

Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213, USA

*Address correspondence to Marcel Adam Just, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA.
Email: just@cmu.edu.

Abstract

The abstractness of concepts is sometimes defined indirectly as lacking concreteness, this view provides little insight into their cognitive or neural basis. Multivariate pattern analytic techniques applied to functional magnetic resonance imaging data were used to characterize the neural representations of 28 individual abstract concepts. A classifier trained on the concepts' neural signatures reliably decoded their neural representations in an independent subset of data for each participant. There was considerable commonality of the neural representations across participants as indicated by the accurate classification of each participant's concepts based on the neural signatures obtained in other participants. Group-level factor analysis revealed 3 semantic 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*. Further exploration of the *Verbal Representation* dimension suggests that the degree a concept is verbally represented can be construed as a point on a continuum between language faculties and perceptual faculties. A predictive model, based on independent behavioral ratings of the 28 concepts along the 3 factor dimensions, provided converging evidence for the interpretations.

Key words: abstract concepts, fMRI, MVPA, predictive modeling, semantic

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 and 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 and 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.

Materials and 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 and Just 2016; Bauer and 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° flip 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 and Just 2016; Bauer and 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 and Just 2016; Bauer and 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 and 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 Classification

A 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).

Between–Participants Classification

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 and 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 and 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, Murphy and Poesio 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 Fig. 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 1).

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

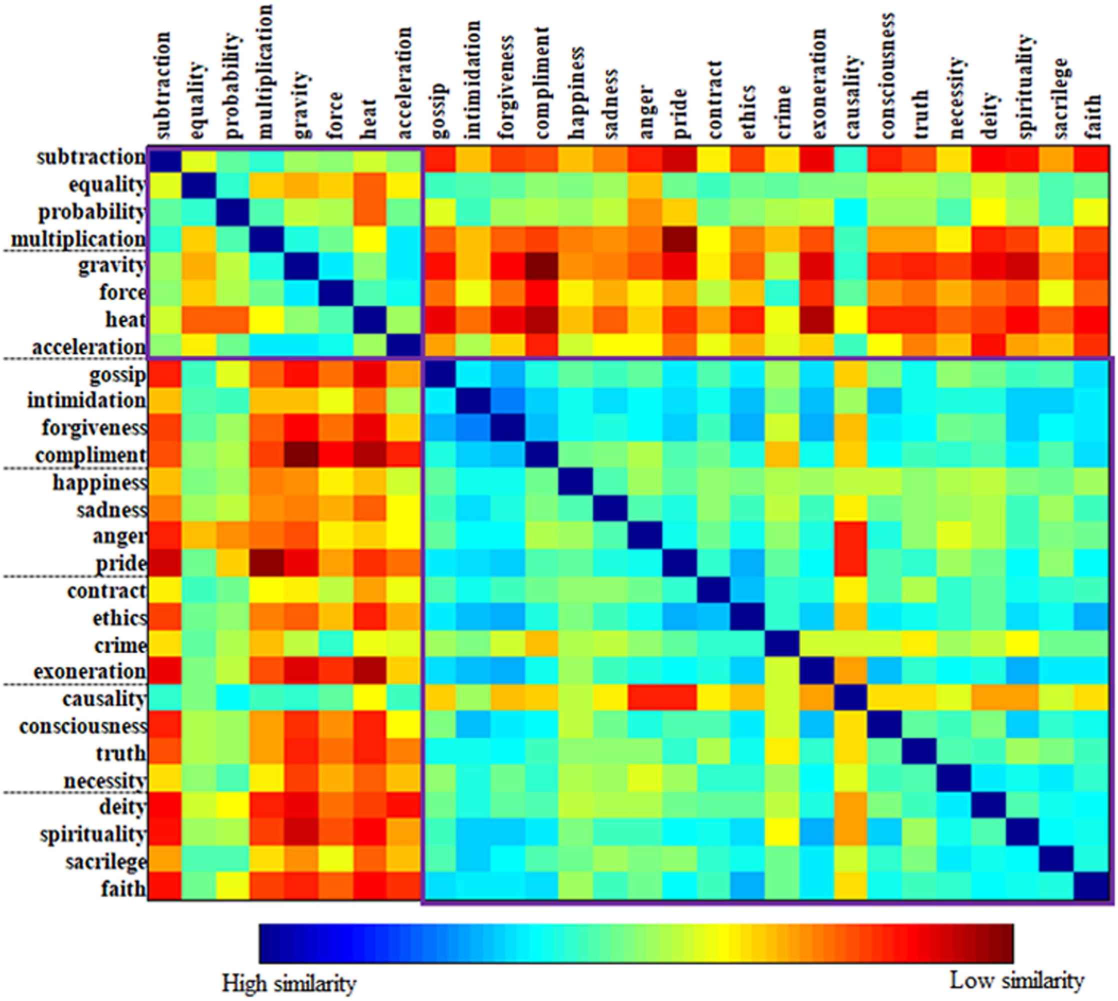

Figure 1. Representational similarity between neural activation (**blue colors indicate higher similarity**) for individual concepts. Dotted lines indicate category separation.

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 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 Fig. 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

Table 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.

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).

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 Fig. 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 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 1). The main cortical region containing voxels with high loadings for this factor was the left posterior cingulate cortex (LPCC; Fig. 2), which is associated with the contextualization of one's self in space and emotions (Maddock et al. 2003; Bird et al. 2015; Guterstam et 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 (Fig. 2). This finding regarding word-length provides a face validity check for the factor analysis methods and interpretations.

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 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) 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

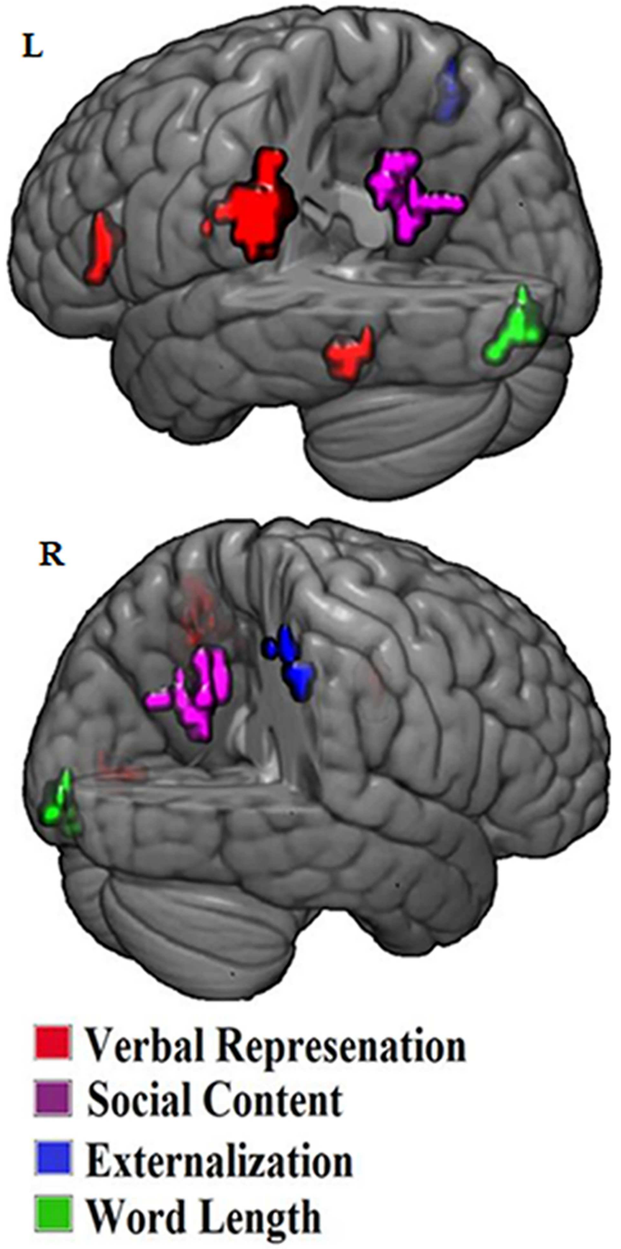

Figure 2. Locations of the voxel clusters with the highest factor loadings for each of the 4 interpretable factors. Voxels were thresholded to have a minimum cluster size of 15 and mean correlations above 0.2 (in either positive or negative direction) between their activation values and their factor 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).

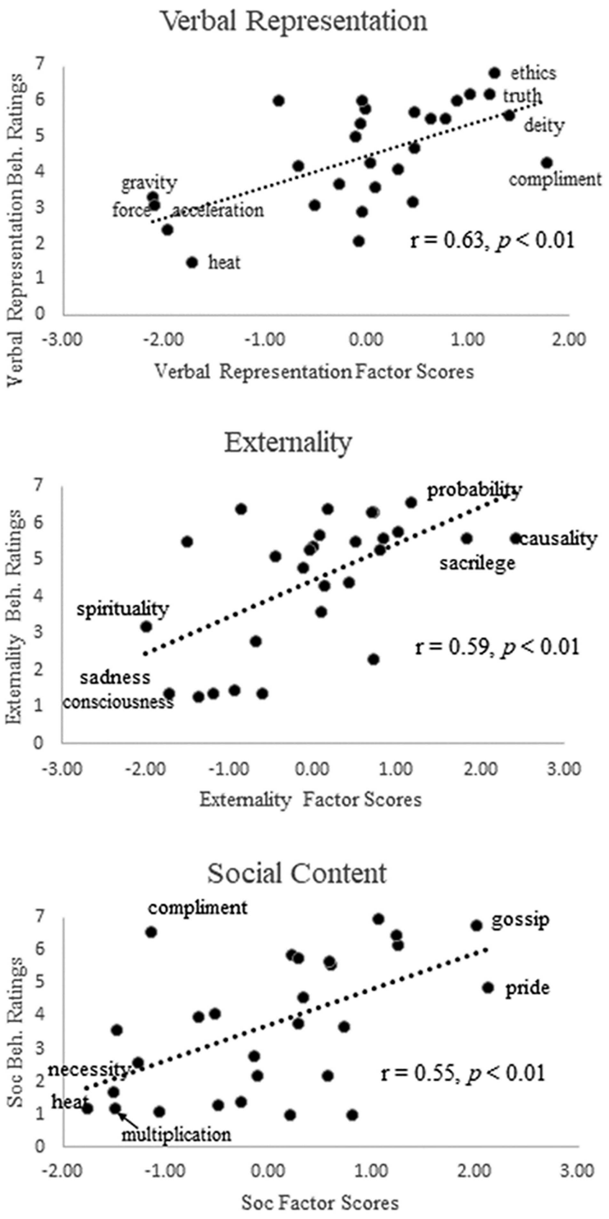

Figure 3. Scatter plot of factor scores for each of the 3 semantic dimensions versus the mean behavioral ratings of the 28 concepts, and their correlations. Although these correlations are significant, they are based on 28 items and therefore the effect sizes should be interpreted with caution.

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

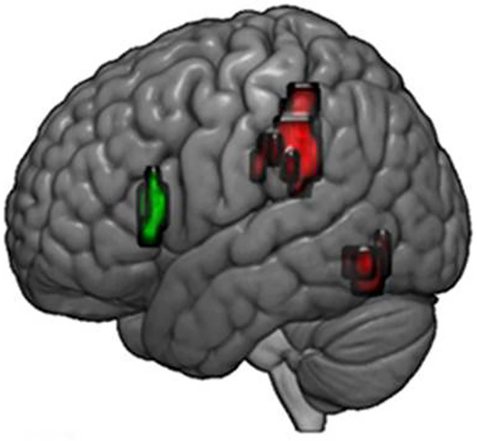

Figure 4. Correlation between the Verbal Representation factor scores and MPSC activation. As concepts become more verbally represented they recruit more LIFG and show less activation in regions associated with the visual representation of concepts (LSMG and LLOC). Positive correlations shown in green; negative correlations in red.

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.

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 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 ($r = -0.47$, $P < 0.05$).

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 and 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 cooccurence 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 and 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 (Fig. 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 a more 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 and Warrington 2005, 2010; Hoffman 2016; Hayes and 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 (Fig. 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.

Funding

Office of Naval Research Grant (N00014-16-1-2694 to M.J.).

Notes

We thank Vlad Cherkassky, Marc Coutanche, Rob Mason, and Tim Verstynen for helpful comments on an earlier version of this paper.

References

Anderson AJ, Murphy B, Poesio M. 2014. Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. *J Cognit Neurosci*. 26(3):658–681.

Barsalou LW. 1999. Perceptual symbol systems. *Behav Brain Sci*. 22(4):577–609.

Barsalou LW. 2003. Abstraction in perceptual symbol systems. *Philos Trans R Soc B: Biol Sci*. 358(1435):1177–1187.

Bauer AJ, Just MA. 2017. A brain-based account of “basic-level” concepts. *NeuroImage*. 161:196–295.

Binder JR, Westbury CF, McKiernan KA, Possing ET, Medler DA. 2005. Distinct brain systems for processing concrete and abstract concepts. *J Cogn Neurosci*. 17(6):905–917.

Bird CM, Keidel JL, Ing LP, Horner AJ, Burgess N. 2015. Consolidation of complex events via reinstatement in posterior cingulate cortex. *J Neurosci*. 35(43):14426–14434.

Brysbaert M, Warriner AB, Kuperman V. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav Res Methods*. 46(3):904–911.

Collins DL, Neelin P, Peters TM, Evans AC. 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *J Comput Assist Tomogr*. 18:192–205.

Coutanche MN, Thompson-Schill SL. 2015. Creating concepts from converging features in human cortex. *Cereb Cortex*. 25(9):2584–2593.

Crutch SJ, Warrington EK. 2005. Abstract and concrete concepts have structurally different representational frameworks. *Brain*. 128:615–627.

Crutch SJ, Warrington EK. 2010. The differential dependence of abstract and concrete words upon associative and similarity-based information: complementary semantic interference and facilitation effects. *Cognit Neuropsychol*. 27(1): 46–71.

Grill-Spector K, Kourtzi Z, Kanwisher N. 2001. The lateral occipital complex and its role in object recognition. *Vision Res*. 41:1409–1422.

Guterstam A, Björnsdotter M, Gentile G, Ehrsson HH. 2015. Posterior cingulate cortex integrates the senses of self-location and body ownership. *Curr Biol*. 25(11):1416–1425.

Hayes JC, Kraemer DJM. 2017. Grounded understanding of abstract concepts: The case of STEM learning. *Cognit Res: Princ Implic*. 2(1):7.

Hoffman P. 2016. The meaning of ‘life’ and other abstract words: insight from neuropsychology. *J Neuropsychol*. 10(2): 317–343.

Jefferies E, Patterson K, Jones RW, Lambon Ralph MA. 2009. Comprehension of concrete and abstract words in semantic dementia. *Neuropsychology*. 23:492–499.

Just MA, Cherkassky VL, Aryal S, Mitchell TM. 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS One*. 5:e8622.

Just MA, Cherkassky VL, Buchweitz A, Keller TA, Mitchell TM. 2014. Identifying autism from neural representations of social interactions: neurocognitive markers of autism. *PLoS One*. 9(12):e113879.

Just MA, Pan L, Cherkassky VL, McMakin D, Cha C, Nock MK, Brent D. 2017. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nat Hum Behav*. 1:911–919.

Kassam KS, Markey AR, Cherkassky VL, Loewenstein G, Just MA. 2013. Identifying emotions on the basis of neural activation. *PLoS One*. 8:e66032.

Maddock RJ, Garrett AS, Buonocore MH. 2003. Posterior cingulate cortex activation by emotional words: fMRI evidence from a valence decision task. *Hum Brain Mapp*. 18(1):30–41.

Mason RA, Just MA. 2016. Neural representations of physics concepts. *Psychol Sci*. 27(6):904–913.

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, Just MA. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*. 320(5880):1191–1195.

Pecher D, Boot I, Van Dantzig S. 2011. Abstract concepts: Sensory-motor grounding, metaphors, and beyond. In: B, Ross, editor. *Advances in Research and Theory (Psychology of Learning and Motivation, Volume 54)*. Burlington (MA): Academic Press. p. 217–248.

Silani G, Lamm C, Ruff CC, Singer T. 2013. Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments. *J Neurosci*. 33(39):15466–15476.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Joliot M. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*. 15(1):273–289.

Vigliocco G, Kousta ST, Della Rosa PA, Vinson DP, Tettamanti M, Devlin JT, Cappa SF. 2014. The neural representation of abstract words: the role of emotion. *Cerebral Cortex*. 24(7):1767–1777.

Wang J, Conder JA, Blitzer DN, Shinkareva SV. 2010. Neural representation of abstract and concrete concepts: a meta-analysis of neuroimaging studies. *Hum Brain Mapp*. 31(10): 1459–1468.

Wang J, Baucom LB, Shinkareva SV. 2013. Decoding abstract and concrete concept representations based on single-trial fMRI data. *Human Brain Mapping*. 34(5):1133–1147.

Wang X, Wu W, Ling Z, Xu Y, Fang Y, Wang X, Binder JR, Men W, Gao J, Bi Y. 2017. Organizational principles of abstract words in the human brain. *Cerebral Cortex*. 28(12):4305–4318.

Yang Y, Wang J, Bailer C, Cherkassky VL, Just MA. 2017. 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 Lang*. 175:77–85.

Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD. 2011. Large-scale automated synthesis of human functional neuroimaging data. *Nat Methods*. 8:665–670.