What Makes Different People's Representations Alike: Neural Similarity Space Solves the Problem of Across-subject fMRI Decoding

Rajeev D. S. Raizada[1] and Andrew C. Connolly[2]

Abstract

■ A central goal in neuroscience is to interpret neural activation and, moreover, to do so in a way that captures universal principles by generalizing across individuals. Recent research in multivoxel pattern-based fMRI analysis has led to considerable success at decoding within individual subjects. However, the goal of being able to decode across subjects is still challenging: It has remained unclear what population-level regularities of neural representation there might be. Here, we present a novel and highly accurate solution to this problem, which decodes across subjects between eight different stimulus conditions. The key to finding this solution was questioning the seemingly obvious idea that neural decoding should work directly on neural activation patterns. On the contrary, to decode across subjects, it is beneficial to abstract away from subject-specific patterns of neural activity and, instead, to operate on the similarity relations between those patterns: Our new approach performs decoding purely within similarity space. These results demonstrate a hitherto unknown population-level regularity in neural representation and also reveal a striking convergence between our empirical findings in fMRI and discussions in the philosophy of mind addressing the problem of conceptual similarity across neural diversity. ■

INTRODUCTION

In cognitive neuroscience, the goal is, in general, not to study the peculiarities of particular individuals' brains but, instead, to find regularities that hold across individuals at the population level. An obstacle to that goal is the fact that different people's brains do not directly match up. They share the same gross anatomy and also share coarse-grained functional distinctions, for example, between animate and inanimate object categories (Martin, 2007; Caramazza & Shelton, 1998; Warrington & Shallice, 1984) such as faces and houses (Epstein & Kanwisher, 1998; Kanwisher, McDermott, & Chun, 1997; McCarthy, Puce, Gore, & Allison, 1997). However, at a finer grain, there are diverse individual differences: For example, the size of V1 in different people can vary by a factor of more than two, and this size variability has perceptual consequences (Duncan & Boynton, 2003). At the level of specific neural representations, pattern recognition algorithms have been used to find the multivoxel neural "fingerprints" elicited by given stimulus conditions (Raizada & Kriegeskorte, 2010; Pereira, Mitchell, & Botvinick, 2009; Haynes & Rees, 2006; Norman, Polyn, Detre, & Haxby, 2006; Haxby et al., 2001). However, just as the literal fingerprints on people's hands are idiosyncratic to individuals, the "neural fingerprints" of representations in their brains may also be subject-unique. Indeed, this has been found to be the case. For example, Shinkareva, Mitchell, and colleagues performed both within- and across-subject decoding and found that "a critical diagnostic portion of the neural representation of the categories and exemplars is still idiosyncratic to individual participants" (Shinkareva et al., 2008).

Whatever commonality there might be between different people's neural representations, it must somehow abstract away from their subject-specific finer-grained neural patterns. Can we find a level of representation that is shared across individuals and that is fine-grained than animate-versus-inanimate but that, unlike subject-specific neural fingerprints, succeeds in capturing across-subject commonalities? A shared level of representation that satisfies these conditions would be able to bridge between different people's neural representational schemes. In other words, it would be able to perform across-subject neural decoding.

A potentially promising candidate level of representation is similarity space, which is the set of pairwise relations between items defined by a similarity measure and which has long served as a powerful tool in psychology for investigating cognitive processing (Edelman, 1998; Medin, Goldstone, & Gentner, 1993; Tversky, 1977; Shepard, 1962). In the neural domain, it has been used for visualizing and comparing overall representational structure (Connolly, Gobbini, & Haxby, 2012; Shinkareva, Malave, Just, & Mitchell, 2011; Kriegeskorte et al., 2008; O'Toole et al., 2007; Hanson, Matsuka, & Haxby, 2004; Edelman, Grill-Spector, Kushnir, & Malach, 1998). However, in seeking

[1]Cornell University, [2]Dartmouth College

to decide whether different people's representational schemes are the same, we need to be able to do more than visualize the broad overall match between them. We need to be able to bridge between the different sets of representations, that is, to perform across-subject neural decoding. However, until now, no method of neural decoding using similarity space has been available.

In all previous work on neural decoding, the inputs to the decoding algorithms have not been similarity values but, instead, have been neural activation values themselves (e.g., Pereira et al., 2009; Haynes & Rees, 2006; Norman et al., 2006; Haxby et al., 2001). However, at a fine grain, these neural activation patterns suffer from the subject-specific idiosyncrasies described above. Across-subject decoding of fine-grained neural representations has therefore remained a challenge.

It might seem almost too obvious to be worth stating that neural decoding should take neural activation patterns as its input. Here, we argue that the seemingly tautological nature of that statement is deceptive. On the contrary, we argue that effective decoding across subjects can be achieved without using neural activation patterns as input. Instead, the similarity relations between those patterns may be used as input rather than the neural patterns themselves. By operating on the similarity relations, the decoding can abstract away from the idiosyncratic and subject-specific nature of the neural activation. To support this claim, we present for the first time a method to perform neural decoding purely within similarity space. We then demonstrate that this method achieves highly accurate across-subject decoding.

METHODS

For the analyses in this article, we used the classic Haxby et al. (2001) data set of object-elicited activation in ventral temporal (VT) cortex, kindly made available on-line by Haxby and the developers of PyMVPA (pymvpa.org/datadb/haxby2001.html). The VT cortex ROIs in that study are included in the on-line data set and were manually traced from anatomical scans to consist of the lingual, parahippocampal, fusiform, and inferior temporal gyri. The neural similarity space for each subject was calculated simply as the spatial correlation between the various stimulus-conditions' activation patterns across VT cortex. The stimulus categories spanned the animate-versus-inanimate distinction but also included a lower level of multiple animate and inanimate subcategories. The animate stimuli were subdivided into cats and faces, and the inanimate stimuli were subdivided into bottles, chairs, houses, scissors, scrambled pictures, and shoes.

We first calculated the VT cortex neural similarities between these eight stimulus conditions for each of the six subjects. The similarity measure was the simplest possible: spatial correlation. Before this pattern correlation step, the voxel time courses were first normalized in intensity by being z-scored, that is, by having their mean values subtracted and being divided by their standard deviations. To avoid normalizing out potentially informative stimulus-evoked signals, these means and standard deviations were calculated from the rest-condition blocks only. Such normalization is standard for pattern-based fMRI analyses (Pereira et al., 2009) and, indeed, for machine-learning studies in general (Han & Kamber, 2006). It is particularly useful for correlation-based analyses, which would otherwise tend to be corrupted by outlier intensity values.

In Figure 1, we present and explain our novel method, simple but highly effective, for performing neural decoding purely within similarity space. Using our new method, which we call "decoding by matching of similarity spaces" or DEMOSS, we show here for the first time that similarity space is indeed able to decode between different people's neural representations. We also show that people's shared representational structure goes beyond the animate-versus-inanimate distinction and extends to the fine-grained level of multiple animate and inanimate subcategories. Moreover, we demonstrate below that, by operating purely within similarity space, this across-subject decoding remains accurate even in the presence of a high degree of neural diversity.

Our analysis code was written in Matlab, and the preprocessing and data extraction were carried out in Python using scripts from PyMVPA (Hanke et al., 2009). To facilitate easy replication and verification of our results, all of the analysis code is provided in the Supplementary Information.

RESULTS

Visualization of Overall Similarity Structure Leaves It Unclear whether Decoding Can Be Achieved

As was remarked in the Introduction, neural similarity space has previously been used for visualizing and comparing overall representational structure by combining it with multidimensional scaling (MDS; Shepard, 1962). Examples of such studies include Connolly et al. (2012), Shinkareva et al. (2011), Kriegeskorte et al. (2008), O'Toole et al. (2007), Hanson et al. (2004), and Edelman et al. (1998). However, until now, no method for using similarity space to perform neural decoding has been available. Given the existence of these visualization studies, it is reasonable to ask whether visualization alone is sufficient to judge whether neural decoding could be performed.

In Figure 2, we show 2-D MDS projections of each individual subject's neural similarity space in the Haxby et al. (2001) data set. Some broad commonalities are readily apparent: Houses and scrambled pictures always stand apart from the other stimuli, and bottles, shoes, and scissors typically cluster together. But it is unclear whether these commonalities are sufficient to allow across-subject decoding. Categories that cluster together in some subjects are quite dissimilar in others (e.g., faces

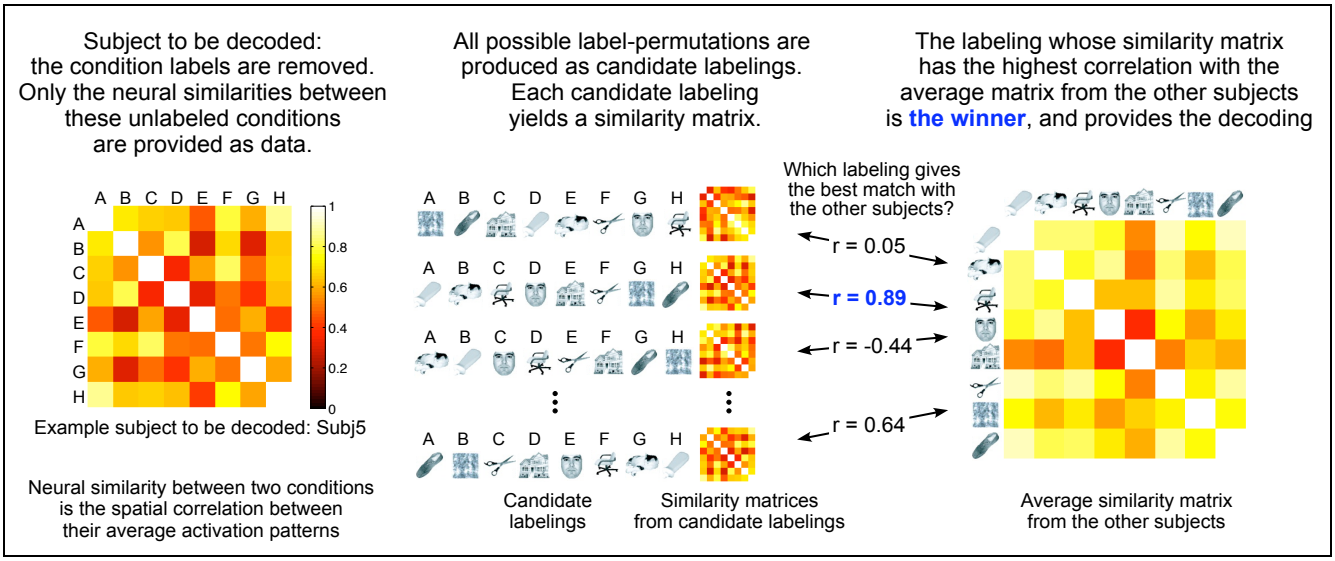

Figure 1. Our novel method of across-subject neural decoding: DEMOSS. The data entered into the model for each subject consists only of the values in their 8 × 8 similarity matrix, constituting 8 × (8 − 1)/2 = 28 unique numbers. Only one permutation-matching computation is performed per subject, so there are no multiple comparisons. The illustrated similarity matrices are the actual data for the example subject shown.

and cats are much more similar in Subjects 1, 2, and 3 than they are in Subjects 4, 5, and 6). Given this, one might expect that a similarity-based decoding would be able to distinguish between faces and cats in the first three subjects but would confuse the two stimulus categories in the remaining three. As we show below, this is not in fact the case: Our similarity-based decoding did not confuse those two categories. The amount of variability across different people's category clusterings means that the visualization, on its own, does not tell us whether an attempt to decode the stimuli across subjects would succeed or fail.

Accurate Across-subject Decoding of Fine-grained Object Categories in VT Cortex

We used our new DEMOSS method, shown in Figure 1, to perform across-subject decoding of the Haxby data. With eight categories per subject and six subjects, there were 48 decodings to perform in all. The method scored 91.7% correct (44 of 48 categories correct). Software to replicate these analyses is provided in the Supplementary Information.

If the animate-versus-inanimate distinction were the level at which different people's neural representational schemes

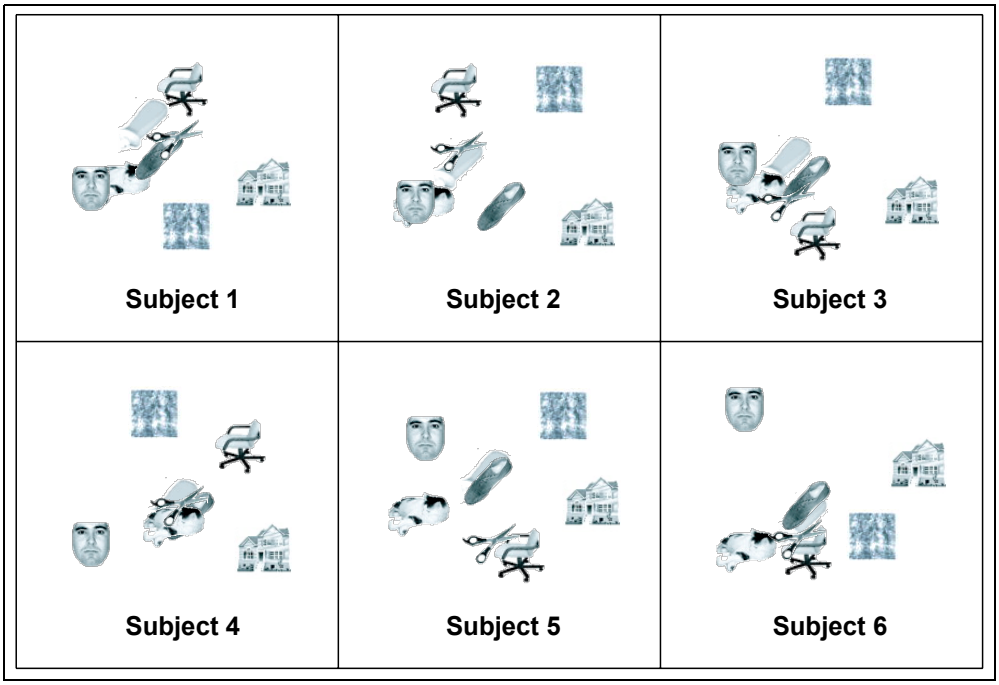

Figure 2. The neural similarity spaces of each of the six subjects in the Haxby et al. (2001) data, visualized in 2-D using MDS (Shepard, 1962). Although some broad commonalities are readily apparent, there are also major intersubject differences. Such a visualization, therefore, leaves it unclear whether similarity space can bridge between different people's neural representations. To directly test that, we need to see whether similarity space enables across-subject decoding.

are the same, then it would be predicted that across-subject decoding should succeed at that level but fail at lower levels in the hierarchy. In contrast, if different people's neural representational schemes are the same not only at the animate-versus-inanimate level but also at lower subdivisions of the hierarchy, then across-subject decoding would be predicted to succeed even at making distinctions between fine-grained subcategories, for example, at distinguishing between different animate categories (faces vs. cats) and between different inanimate categories (e.g., bottles vs. shoes).

The latter prediction held true: The decoding was highly accurate at distinguishing between fine-grained animate and inanimate subcategories. Within the animate subdivision, decoding was 100% correct: A face was never confused with a cat. The more difficult decoding task was within the inanimate subdivision, in which some errors were made; five of the six subjects had all six of their inanimate categories perfectly decoded, and the remaining subject had two pairs of confusions: bottle–scissors and shoe–chair. However, decoding between inanimate subcategories was far above chance (32 of 36 correct decodings, i.e., 88.9% correct). Chance-level performance is to get one eighth of the decodings correct, that is, 12.5%.

Decoding Remains Accurate Even across Widespread Neural Diversity

The success of this across-subject decoding shows that neural similarity space captures a representational scheme that is shared across individuals, even at the fine-grained level of multiple animate and inanimate subcategories. However, as was noted in the Introduction, one of the main sources of difficulty for across-subject neural decoding is the fact that different people's brains do not directly match up. In the analysis above, that difficulty was not felt with its full force because all of the neural signals were drawn from the same brain area: VT cortex.

As Figures 3A and C shows, VT cortex masks drawn by Haxby et al. on individual subjects' anatomical scans do not completely overlap when they are aligned to a common space, but they do mostly overlap. A stronger test would, therefore, be to use anatomically dispersed and highly variable sets of voxels in different individuals.

To carry that out, we devised a simple feature-selection scheme to find informative voxels within each individual subject. For each voxel, we calculated two measures to be used for selection. Comparing all of the visual object stimuli together against the rest blocks, we determined the t statistic for the degree to which each voxel was active. Then, considering only the object stimuli blocks, we calculated the F statistic of the ratio of between-class variance to within-class variance. We then selected the voxels within each subject that scored not only in the top 5% of t values but also in the top 5% of F values, that is, the voxels that were active and that differentiated between the various object stimuli. As before, the neural similarity space for each subject was calculated simply as the spatial correlation between the various stimulus-conditions'

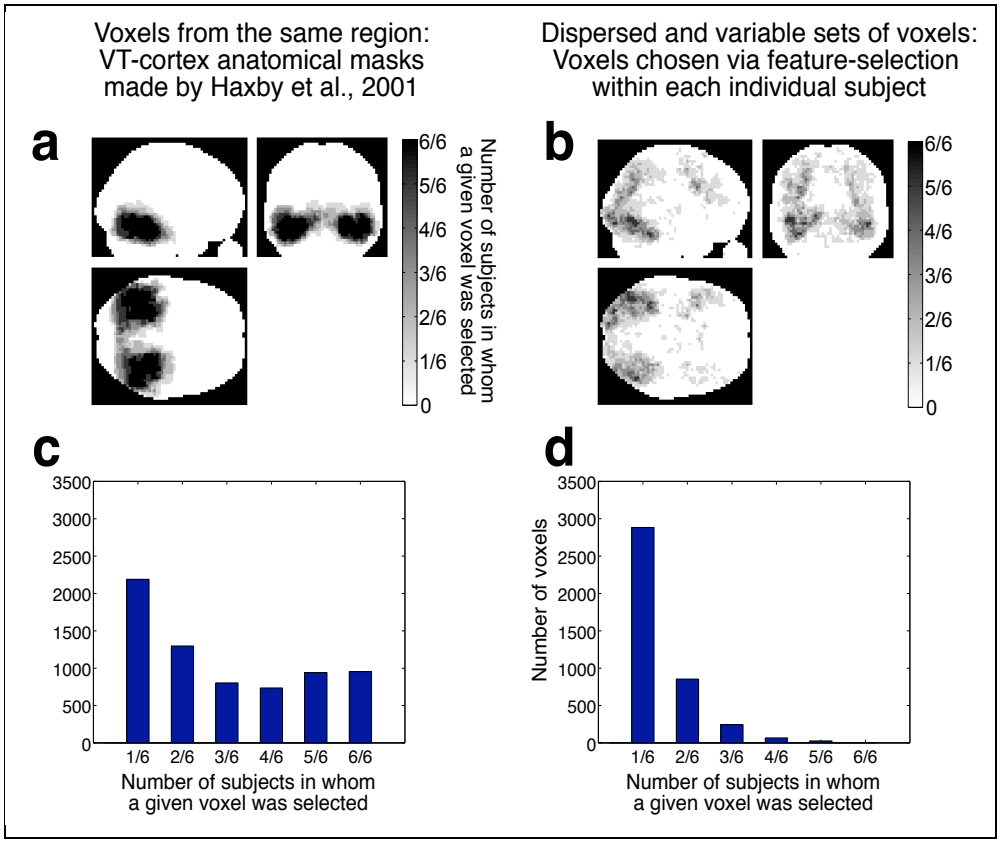

Figure 3. Maximum intensity projections and histograms showing differing degrees of across-subject neural diversity. (A, C) When the voxels used for across-subject decoding were specified by the VT cortex masks included in the Haxby data set, the performance was 91.7% correct. However, there was relatively little neural diversity across different people's VT masks. (B, D) In a separate analysis, we used a simple feature-selection scheme to find different sets of informative voxels within each individual subject. The selected voxels were anatomically dispersed and highly variable across different individuals. Nonetheless, using these diverse sets of voxels, the across-subject decoding still achieved 87.5% correct. Chance-level performance is 12.5%.

activation patterns, but this time, the patterns were the activations across the selected voxels rather than across the VT cortex region. (Matlab scripts used to perform this feature selection and to carry out the similarity analyses on the selected voxels are provided in the Supplementary Information). To compare the locations of the selected voxels across different subjects, the brain volumes were all spatially normalized to the standard MNI-152 template at $3 \times 3 \times 3$ mm resolution using SPM8, before feature selection or similarity analysis was carried out.

These feature-selected voxels showed a very high degree of diversity across subjects: The number of voxels selected within each subject ranged from 473 to 1346. In other words, the dimensionalities of people's neural activation spaces varied widely across different individuals. It is unclear even how to compare a 473-dimensional space with a 1346-dimensional space, let alone to try to decode between them. However, by calculating the spatial correlations between the stimulus-elicited activation patterns within each activation space, the different subjects' activation spaces, with their widely varying dimensionalities, all become transformed into eight-dimensional similarity spaces defined by the eight stimulus categories. These similarity spaces can be compared, and using our novel DEMOSS method presented in Figure 1, we can perform across-subject decoding between them.

Different people's selected voxels varied not only in their number but also in their locations across the subjects' brains. As Figures 3B and D shows, the selected voxels were dispersed broadly throughout the brain, and their anatomical locations were highly variable across subjects. As would be expected, the greatest concentration of selected voxels was found in VT cortex; however, informative voxels were found in many other regions, including parietal and frontal cortex. The selected voxel locations in those areas were often shared by just one or two subjects, as can be seen from the light-gray regions in the maximum intensity projection in Figure 3B. The histogram in Figure 3D confirms that the majority of selected voxel locations were specific to individual subjects and that very few voxel locations were shared by multiple individuals. As it happened, there was not a single voxel that was selected in all six subjects, not even in the heart of VT cortex.

To what degree would neural decoding be able to succeed in the face of this very marked neural diversity? Using the similarity spaces derived from these disparate sets of feature-selected voxels, the performance of across-subject decoding was 87.5%, only slightly lower than the 91.7% obtained when the voxels were specified by VT cortex masks. Four of the six subjects were decoded perfectly. In one subject, there were four confusions: cat, chair, face, and scissors. In the remaining subject, bottles and scissors were confused. As before, chance-level performance is 12.5%. Thus, similarity space was able to capture the representational scheme shared across individuals, although the neural populations used to match people's representations were extremely diverse.

Decoding Works Best in VT Cortex Compared with Other Areas

In the analyses in the preceding sections, we have shown two results: First, accurate across-subject decoding can be achieved by matching neural similarities derived from an anatomically defined VT cortex ROI. Those VT cortex voxels are shown in Figure 3A. Second, decoding remains accurate even in the face of widespread neural diversity, that is, when each subject's neural similarities are derived from subject-specific sets of selected voxels, shown in Figure 3B.

However, those two results do not address the following question: Are there other brain areas, apart from VT cortex, that also allow successful across-subject decoding? Might it perhaps be the case that the decoding's success does not actually arise because of population-level regularities in object recognition representations? For example, if regions that are believed not to participate in visual object recognition, such as auditory or somatosensory cortex, were found to produce neural similarities yielding accurate across-subject decoding, then some quite different interpretation of our results would be required.

To test this, we performed our decoding using a full set of cortical anatomical ROIs from the standard Harvard–Oxford atlas (Desikan et al., 2006), which is distributed with the FMRIB Software Library fMRI analysis package (Smith et al., 2004). There are 48 bilateral cortical regions in that atlas. Using voxels from each of those regions in turn, we calculated neural similarities for all of the subjects and applied our similarity-based across-subject decoding, yielding an overall percentage-correct score for each region. The results are shown in Figure 4. Decoding accuracies ranged from 0% in the frontal operculum to 77.1% in the temporal occipital fusiform cortex. The inferior lateral occipital cortex also scored well at 70.8% correct. As Figure 4 shows, the ventral visual stream contained the most accurately decoding regions, corroborating the hypothesis that it is indeed visual object representations that are driving the decoding's success.

The 77.1% score in temporal occipital fusiform cortex shows that this region contains a great deal of information about visual object representations and that the similarity structure of those representations is highly conserved across subjects. However, that score is markedly worse than the accuracies obtained from spatially larger and more distributed samples of cortical tissue, either the entire VT cortex region (shown in Figure 3A and yielding 91.7% correct) or the diverse sets of voxels spread across VT, parietal, and frontal regions (shown in Figure 3B and yielding 87.5% correct). Thus, although the fusiform region within VT cortex contains more robust object representations than any other individual ROI, the representations distributed across broader expanses of cortex are still stronger.

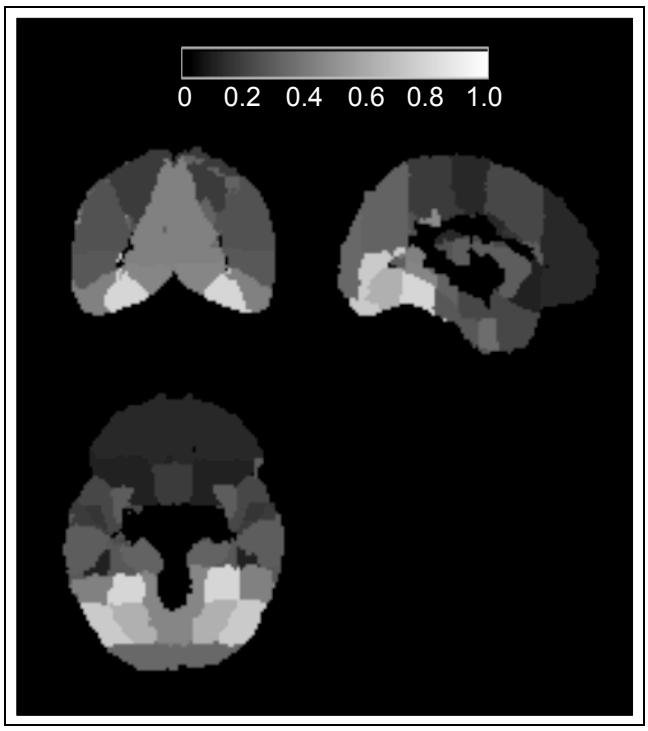

Figure 4. Across-subject decoding accuracies obtained from each individual cortical ROI within the standard Harvard–Oxford anatomical atlas (Desikan et al., 2006). It can be seen that ventral visual stream contained the most accurately decoding regions, corroborating the hypothesis that it is indeed visual object representations that are driving the decoding's success.

Statistical Significance Tests for a Permutation Distribution

Given that there are eight stimulus categories, the number of possible category labelings for each subject's decoding is equal to the factorial of eight, that is, 40,320. Only one of those more than 40,000 labelings gets all eight out of eight labelings correct. It is therefore noteworthy that this solitary perfect eight-out-of-eight labeling often emerged as the decoding output, in virtue of its having a higher level of across-subject match in similarity space than any of the other 40,000 labelings. That eight-out-of-eight perfect decoding was achieved for five of the six subjects when all of the neural information was drawn from cortical area VT, and for four of the six subjects when the neural information was drawn from anatomically dispersed and highly variable feature-selected sets of voxels. For the subjects whose decoding was not perfect, it was still significantly above chance. As the Supplementary Information and its accompanying Matlab code show, the chance level is to get one of eight correct and to meet a significance of $p < .05$, three or more of the eight categories need to be decoded correctly. In our results, not only when decoding from VT cortex but also when using the feature-selected voxels, no subject's decoding achieved fewer than four of eight correct. This suggests that similarity space, even with its very simple construction and greatly reduced dimensionality, does indeed succeed in capturing a crucial aspect of what different people's representations have in common.

DISCUSSION

The results above demonstrate for the first time, using real neural data, that similarity space can provide a decoding between different people's representational schemes. This across-subject decoding remains highly accurate even when the neural similarities are derived from widely diverse sets of voxels across different subjects.

Relation to Previous fMRI Studies

Previous across-subject decoding attempts have all involved feeding thousands of voxels and hundreds of time points into classifier algorithms, so it has remained unclear which aspects of the complex neural signal have been the ones that different people shared. In contrast, our new across-subject decoding method takes in an extremely reduced data set as input: only the similarity space of people's neural category representations. The success of its decoding is therefore driven entirely by across-subject commonalities in that abstract category-similarity structure.

Moving from the nature of the input to questions of performance, our study decodes fine-grained category distinctions across subjects with high accuracy, whereas previous studies have either decoded coarse-grained changes in brain state or with lower accuracy or both. An example of decoding a large-scale change in brain state is distinguishing between the performance of different behavioral tasks, such as reading a sentence versus looking at a picture (Wang, Hutchinson, & Mitchell, 2003) and face matching versus location matching (Mourao-Miranda, Bokde, Born, Hampel, & Stetter, 2005), or between several different cognitive tasks (Poldrack, Halchenko, & Hanson, 2009). A different example of a coarse-grained distinction is between being rewarded with money versus viewing an attractive face (Clithero, Smith, Carter, & Huettel, 2011). Shinkareva, Mitchell, and colleagues (Shinkareva et al., 2008) went further and were able to decode not only which general category of object a person was looking at (tool vs. dwelling) but also which of the five specific exemplars within each category they were looking at. However, their across-subject decoding, which operated directly on neural activation, worked for only 8 of their 12 subjects and achieved a considerably lower level of performance than our approach, which operates instead on neural similarities. Another interesting and important line of work in this area is that of Haxby et al. (2011), who have proposed a high-dimensional mapping called "hyper-alignment" of one person's voxel space onto another's.

Following its initial publication (Haxby et al., 2001), there have been a number of subsequent articles containing analyses of the Haxby et al. (2001) data. These prior

studies have performed within-subject voxel-wise sensitivity analyses (Hanson et al., 2004), compared the performance of classifiers applied to the image stimuli themselves against that of classifiers applied to the neural data (O'Toole et al., 2007; O'Toole, Jiang, Abdi, & Haxby, 2005), investigated ICA (Daubechies et al., 2009), and explored the use of classifier ensembles (Kuncheva, Rodriguez, Plumpton, Linden, & Johnston, 2010). Two of these Haxby data studies (O'Toole et al., 2007; Hanson et al., 2004) used similarity structure analyses to explore representational organization. However, neither those studies nor any previous investigations of neural similarity space have used similarity space to perform across-subject decoding. This across-subject decoding is a key contribution of our new approach, along with what it tells us about people's shared hierarchy of object representations and its ability to find the same representations across highly diverse neural populations. Our new approach also intersects with some longstanding conceptual debates in cognitive science, as we discuss in the following section.

Relation to Longstanding Conceptual Debates in Cognitive Science

A striking aspect of the solution to across-subject decoding presented above is its parallelism to a proposal made more than 20 years ago by the neuro-philosopher Paul Churchland (Churchland, 1986). Churchland proposed, on purely theoretical grounds, that matched structure in people's neural similarity spaces could explain how different brains can form the same mental representations. He referred to this as "the problem of conceptual similarity across neural diversity" (Churchland, 1998).

However, that proposal has faced opposition, most notably from Fodor and Lepore (1992, 1999) who argued that similarity space theories cannot explain how different people could possess the same concept. Partly in response to such objections, researchers have presented computer simulations as evidence that similarity space could indeed, in principle, provide a solution (Goldstone & Rogosky, 2002; Laakso & Cottrell, 2000). However, no simulation can address the question of how real brains actually do solve the problem. Our results above, using real neural data, do precisely that.

The question of whether the philosophy of mind is a relevant part of cognitive science is far beyond the scope of this article. We merely remark that when completely different lines of inquiry, originating respectively from conceptual and empirical concerns, end up converging on the same solution, it may be an indication that they are both being guided by something real.

How Generalizable Will Similarity-based Decoding Turn Out to Be?

How robust and generalizable our proposed solution to across-subject decoding will turn out to be is, of course, an empirical question. This article presents its successful decoding of the classic and much-studied data set from Haxby et al. (2001), but clearly multiple diverse data sets will need to be analyzed in order for its generalizability to be determined. By presenting the source code for our analyses (in Matlab and Python) in the Supplementary Information, we hope to facilitate such tests.

Some preliminary evidence for generalizability comes from two of our other studies: One, still in progress, and the other, completed and under review. It is beyond the scope of the present article to describe those studies in full, but here, we provide a brief outline. The first, currently under review, applies our similarity-based decoding approach to the problem of decoding the meanings of words using the publicly available data set from the study by Mitchell et al. (2008). Our decoding of that data is based on the simple hypothesis that neural similarity matches semantic similarity. It achieves more accurate decoding of untrained word pairs than was obtained in the original Mitchell et al. study or in any other published analysis of their data. For more details, please see Raizada (under review).

Probably, the most impressive demonstration to date of the power of neural similarity is the across-species study by Kriegeskorte, Kiani, and colleagues (Kriegeskorte et al., 2008; Kiani, Esteky, Mirpour, & Tanaka, 2007), who demonstrated a strikingly high degree of match between the neural similarity structures of human and monkey inferotemporal cortex. An interesting question is whether the degree of match between those similarity structures is sufficient to enable across-species decoding. Our new similarity-based decoding allows that question to be directly addressed, and we are currently collaborating with Kriegeskorte and Kiani to apply our method to their data. Their data use 92 stimulus categories, as opposed to the eight categories in the Haxby data. Whereas an exhaustive search through 8! (8 factorial, i.e., 40,320) possible labelings takes only a few seconds on a standard desktop computer, an exhaustive search through 92! labelings would be computationally intractable. In collaboration with Kriegeskorte and Kiani, we have developed a simple heuristic for searching through labelings that is fast and quite effective, and enables across-species decoding of their data with accuracies much greater than chance. An article describing this new work is currently in preparation (Raizada, Kiani, & Kriegeskorte, in preparation).

Limitations of the New Approach

Although the new studies described above provide preliminary evidence for the generality of our new decoding approach, there are some types of decoding to which it would not be applicable.

The simplest and probably the most common scenario would be when there are only two categories to be decoded. To see why our decoding approach would not handle such decoding, let us call them Category 1 and Category 2,

and let the task of the decoding be to assign the labels A and B to those two categories. The matrix of neural similarities between the two categories would therefore consist of four elements. As is the case for all similarity matrices, the diagonal elements would have the value of one, with each diagonal element being the neural similarity of a given category with itself. The more interesting elements are the off-diagonal entries: One is the similarity of A with B, and the other is the similarity of B with A. For any well-defined similarity metric, those two values will be equal. Therefore, a candidate decoding that assigns the label A to Category 1 and the label B to Category 2 will produce the same similarity matrix as a decoding whose labeling is the reverse. Our decoding approach works by selecting the similarity matrix with the highest degree of across-subject match, but in the two-category case, both degrees of match will necessarily be equal.

It may be helpful, at this point, to distinguish between two senses of symmetry in a similarity matrix. Because the values of similarity(A,B) and similarity(B,A) are equal, all similarity matrices are symmetric about the axis of their leading diagonal, that is, they are equal to their own matrix transpose. That standard type of symmetry is unproblematic for our approach. However, the similarity matrix produced by the two-category case described above is symmetric in an additional sense: Different permutations of labels end up producing the same similarity matrix. We will refer to this as "permutation symmetry." The two labeling permutations (A = 1, B = 2) and (B = 1, A = 2) are permutation-symmetric, and indeed, any two-category similarity matrix will be permutation-symmetric.

It is also possible for permutation symmetry to arise even when there are more than two categories. For example, if there are three categories whose pairwise neural similarities define an equilateral triangle or four categories whose similarities define an equilateral tetrahedron, then the resulting similarity matrices would be entirely permutation-symmetric, and our decoding approach would be unable to deal with them.

In summary, our method is limited to sets of stimuli that are not permutation-symmetric. Thus, it cannot be applied to two-category sets. When there are more than two categories, only specially constructed stimulus sets such as those described above will have the property of permutation symmetry. In most stimulus sets, such as those of Kriegeskorte et al. (2008), Mitchell et al. (2008), and Haxby et al. (2001), the stimuli were not constructed to have identical similarity relations with each other. They are, therefore, not permutation-symmetric, and our approach decodes those stimuli with success.

Two more minor limitations should also be borne in mind. First, our approach is designed for across-subject decoding, and indeed, it achieves that decoding by abstracting away from individual subjects' neural activation patterns. It is therefore not applicable to single-subject decoding. Second, like all existing work on across-subject decoding, our approach requires that the same set of stimulus categories must be used for all of the subjects. This current limitation raises some interesting possibilities: If different subjects are presented with partially but incompletely overlapping sets of stimuli, then it might be possible to use a model of the stimulus space (Kay, Naselaris, Prenger, & Gallant, 2008; Mitchell et al., 2008) to interpolate across those partially overlapping sets. That possibility is outside the scope of the current article, but we plan to explore it in future work.

Conclusion

It might seem obvious, at first sight, that neural decoding should take neural activation patterns as its input. Indeed, all previous neural decoding approaches have done exactly that; this article is, as far as we are aware, the first to perform neural decoding using not the neural patterns themselves but, instead, the similarities between those patterns. This, we wish to argue, is precisely why it is able to achieve accurate across-subject decoding. To capture the commonalities across subjects, it is beneficial to abstract away from their idiosyncratic and subject-specific "neural fingerprints." Performing the decoding in similarity space does exactly that.

The concept of similarity has been found to be a powerful tool in multiple domains of cognitive psychology (Edelman, 1998; Medin et al., 1993; Tversky, 1977; Shepard, 1962) and in studies of language and conceptual structure (Storms, Navarro, & Lee, 2010; Pedersen, Patwardhan, & Michelizzi, 2004; Landauer & Dumais, 1997; Miller, 1995). The idea of classifying stimuli based on their similarities, rather than on the features of the stimuli themselves, has also attracted considerable attention in the machine-learning literature (Chen, Garcia, Gupta, Rahimi, & Cazzanti, 2009; Pekalska & Duin, 2005). In fMRI research, most investigations of neural similarity have been in the domain of visual object recognition (Connolly et al., 2012; Shinkareva et al., 2011; Kriegeskorte et al., 2008; O'Toole et al., 2007; Hanson et al., 2004; Edelman et al., 1998), but it has also been found to be important in memory (Xue et al., 2010) and olfaction (Howard, Plailly, Grueschow, Haynes, & Gottfried, 2009). Indeed, in animal neurophysiology studies of olfaction, the concept of neural similarity is central (Dupuy, Josens, Giurfa, & Sandoz, 2010; Haddad et al., 2008; Guerrieri, Schubert, Sandoz, & Giurfa, 2005; Cleland, Morse, Yue, & Linster, 2002). These considerations suggest that our new approach for decoding in similarity space may have broad applicability, across multiple neural and behavioral domains.

Our across-subject neural decoding demonstrates the match between different people's representational schemes by accurately bridging between them. It achieves this by operating entirely within similarity space. Whether drawing upon neural information from within a specific cortical area or from disparate and diverse neural populations, this reveals a population-level regularity that makes different people alike.

Acknowledgments

We thank Jim Haxby for permission to use the data set from his 2001 *Science* article and Daniel Ansari, Silvia Bunge, Shimon Edelman, Niko Kriegeskorte, and Russ Poldrack for very helpful comments on earlier versions of the manuscript. Andrew Connolly was funded by NIMH NRSA Grant 1F32MH085433-01A1.

Reprint requests should be sent to Rajeev D. S. Raizada, Department of Human Development, Cornell University, Martha Van Rensselaer Hall, Ithaca, NY 14853, or via e-mail: raizada@cornell.edu.

REFERENCES

Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain the animate–inanimate distinction. *Journal of Cognitive Neuroscience, 10,* 1–34.

Chen, Y., Garcia, E., Gupta, M., Rahimi, A., & Cazzanti, L. (2009). Similarity-based classification: Concepts and algorithms. *The Journal of Machine Learning Research, 10,* 747–776.

Churchland, P. M. (1986). Some reductive strategies in cognitive neurobiology. *Mind, 95,* 279–309.

Churchland, P. M. (1998). Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *Journal of Philosophy, 95,* 5–32.

Cleland, T. A., Morse, A., Yue, E. L., & Linster, C. (2002). Behavioral models of odor similarity. *Behavioral Neuroscience, 116,* 222–231.

Clithero, J. A., Smith, D. V., Carter, R. M., & Huettel, S. A. (2011). Within- and cross-participant classifiers reveal different neural coding of information. *Neuroimage, 56,* 699–708.

Connolly, A. C., Gobbini, M. I., & Haxby, J. V. (2012). Three virtues of similarity-based multivariate pattern analysis: An example from the human object vision pathway. In N. Kriegeskorte & G. Kreiman (Eds.), *Understanding visual population codes: Toward a common multivariate framework for cell recording and functional imaging*. Cambridge, MA: MIT Press.

Daubechies, I., Roussos, E., Takerkart, S., Benharrosh, M., Golden, C., D'Ardenne, K., et al. (2009). Independent component analysis for brain fMRI does not select for independence. *Proceedings of the National Academy of Sciences, U.S.A., 106,* 10415–10422.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage, 31,* 968–980.

Duncan, R. O., & Boynton, G. M. (2003). Cortical magnification within human primary visual cortex correlates with acuity thresholds. *Neuron, 38,* 659–671.

Dupuy, F., Josens, R., Giurfa, M., & Sandoz, J.-C. (2010). Calcium imaging in the ant camponotus fellah reveals a conserved odour-similarity space in insects and mammals. *BMC Neuroscience, 11,* 28.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences, 21,* 449–498.

Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology, 26,* 309–321.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392,* 598–601.

Fodor, J., & Lepore, E. (1999). All at sea in semantic space: Churchland on meaning similarity. *Journal of Philosophy, 96,* 381–403.

Fodor, J. A., & Lepore, E. (1992). *Holism: A shopper's guide*. Oxford, U.K.: Blackwell.

Goldstone, R. L., & Rogosky, B. J. (2002). Using relations within conceptual systems to translate across conceptual systems. *Cognition, 84,* 295–320.

Guerrieri, F., Schubert, M., Sandoz, J.-C., & Giurfa, M. (2005). Perceptual and neural olfactory similarity in honeybees. *PLoS Biology, 3,* e60.

Haddad, R., Khan, R., Takahashi, Y. K., Mori, K., Harel, D., & Sobel, N. (2008). A metric for odorant comparison. *Nature Methods, 5,* 425–429.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). Amsterdam: Elsevier.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Olivetti, E., Fründ, I., Rieger, J. W., et al. (2009). PyMVPA: A unifying approach to the analysis of neuroscientific data. *Frontiers in Neuroinformatics, 3,* 3.

Hanson, S. J., Matsuka, T., & Haxby, J. V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: Is there a "face" area? *Neuroimage, 23,* 156–166.

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,* 2425–2430.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., et al. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron, 72,* 404–416.

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

Howard, J. D., Plailly, J., Grueschow, M., Haynes, J.-D., & Gottfried, J. A. (2009). Odor quality coding and categorization in human posterior piriform cortex. *Nature Neuroscience, 12,* 932–938.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17,* 4302–4311.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature, 452,* 352–355.

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology, 97,* 4296–4309.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60,* 1126–1141.

Kuncheva, L. I., Rodriguez, J. J., Plumpton, C. O., Linden, D. E. J., & Johnston, S. J. (2010). Random subspace ensembles for fMRI classification. *IEEE Transactions on Medical Imaging ing, 29,* 531–542.

Laakso, A., & Cottrell, G. (2000). Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology, 13,* 47–76.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology, 58,* 25–45.

McCarthy, G., Puce, A., Gore, J., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience, 9,* 605–610.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254–278.
Miller, G. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38,* 39–41.
Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320,* 1191–1195.
Mourao-Miranda, J., Bokde, A. L. W., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. *Neuroimage, 28,* 980–995.
Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences, 10,* 424–430.
O'Toole, A. J., Jiang, F., Abdi, H., & Haxby, J. V. (2005). Partially distributed representations of objects and faces in ventral temporal cortex. *Journal of Cognitive Neuroscience, 17,* 580–590.
O'Toole, A. J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J. P., & Parent, M. A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of Cognitive Neuroscience, 19,* 1735–1752.
Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). Wordnet:: Similarity: Measuring the relatedness of concepts. In S. Dumais, D. Marcu, & S. Roukos (Eds.), *Demonstration papers at HLT-NAACL 2004 on XX* (pp. 38–41). Boston, MA: Association for Computational Linguistics.
Pekalska, E., & Duin, R. P. W. (2005). *The dissimilarity representation for pattern recognition: Foundations and applications*. Hackensack, NJ: World Scientific.
Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage, 45(Suppl. 1),* S199–S209.
Poldrack, R. A., Halchenko, Y. O., & Hanson, S. J. (2009). Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science, 20,* 1364–1372.
Raizada, R. D. S. (under review). Representing the meanings of words: Neural similarity matches semantic similarity.
Raizada, R. D. S., Kiani, R., & Kriegeskorte, N. (in preparation). Across-species neural decoding reveals the similarities and differences in how humans and monkeys represent the visual world.
Raizada, R. D. S., & Kriegeskorte, N. (2010). Pattern-information fMRI: New questions which it opens up, and challenges which face it. *International Journal of Imaging Systems and Technology, 20,* 31–41.
Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27,* 125–140.
Shinkareva, S. V., Malave, V. L., Just, M. A., & Mitchell, T. M. (2011). Exploring commonalities across participants in the neural representation of objects. *Human Brain Mapping*. doi: 10.1002/hbm21296.
Shinkareva, S. V., Mason, R. A., Malave, V. L., Wang, W., Mitchell, T. M., & Just, M. A. (2008). Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS ONE, 3,* e1394.
Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage, 23(Suppl. 1),* S208–S219.
Storms, G., Navarro, D. J., & Lee, M. D. (2010). Introduction to the special issue on formal modeling of semantic concepts. *Acta Psychologica (Amsterdam), 133,* 213–215.
Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327.
Wang, X., Hutchinson, R., & Mitchell, T. (2003). Training fMRI classifiers to detect cognitive states across multiple human subjects. In S. Thrun (Ed.), *Proceedings of the 2003 Conference on Neural Information Processing Systems,* Vancouver.
Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain, 107,* 829–854.
Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater neural pattern similarity across repetitions is associated with better memory. *Science, 330,* 97–101.