Inductive inference based on probability and similarity*

Matthew Weber
Princeton University

Daniel Osherson
Princeton University

August 21, 2008

Abstract

We advance a theory of inductive inference designed to predict the conditional probability that certain natural categories satisfy a given predicate given that others do (or do not). A key component of the theory is the similarity of the categories to one another. We measure such similarities in terms of the overlap of metabolic activity in voxels of various posterior regions of the brain in response to viewing instances of the category. The theory and similarity measure are tested against averaged probability judgments elicited from a separate group of subjects. Fruit serve as categories in the present experiment; results are compared to earlier work with mammals.

1 Introduction

The quality of an inductive inference from premises $\mathsf{P}_1 \cdots \mathsf{P}_n$ to conclusion C can be conceptualized in alternative ways. It may be understood as the strength of the causal connection between believing $\mathsf{P}_1 \cdots \mathsf{P}_n$ and believing C (Osherson et al., 1990), as the degree to which $\mathsf{P}_1 \cdots \mathsf{P}_n$ confirms C in the Bayesian sense (Tentori et al., 2007), or as

*Weber acknowledges support from an NSF graduate research fellowship. Osherson acknowledges the Henry Luce Foundation. Contact: matthew w/osherson@princeton.edu.

the subjective conditional probability of C given $P_1 \cdots P_n$. Here we adopt the latter perspective, and discuss a theory of the conditional probability $Prob(C \mid P_1 \cdots P_n)$ that people assign to C on the supposition that $P_1 \cdots P_n$. Our theory is limited to the special case in which all the statements $C, P_1 \cdots P_n$ have subject-predicate form Qa for a common predicate Q applied to natural categories a. To illustrate, the categories might be fruit, and the predicate: *are a significant source of at least* 10 *dietary vitamins*. Then, our theory bears on answers to questions like the following.

(1) (a) What is the probability that peaches are a significant source of at least 10 dietary vitamins assuming that this is true for plums?

(b) What is the probability that oranges are a significant source of at least 10 dietary vitamins assuming that this is NOT true for strawberries?

(c) What is the probability that pears are a significant source of at least 10 dietary vitamins assuming that this is true for apples and avocados?

(d) What is the probability that tomatoes are a significant source of at least 10 dietary vitamins assuming that this is true for grapes but NOT bananas?

More generally, for natural categories like fruit we consider predicates of a biological character about which typical reasoners have partial but not definite knowledge; in particular, the predicates are not "blank" like *possesses trace quantities of molybdenum*.

To predict responses to queries like (1), we describe a function with two kinds of inputs and the desired conditional probabilities as outputs. One input is the set of *unconditional* probabilities relevant to a given inference, e.g., for (1)b, the reasoner's estimates of:

(2) (a) The probability that oranges are a significant source of at least 10 dietary vitamins.

(b) The probability that strawberries are a significant source of at least 10 dietary vitamins.

The other input is the perceived similarity among the categories figuring in the argument, e.g., for (1)c, the similarity between pears and apples, between pears and avocados, and between apples and avocados. The similarity between categories a and b will

be denoted $sim(\mathfrak{a}, \mathfrak{b})$ and assumed to satisfy $sim(\mathfrak{a}, \mathfrak{b}) = sim(\mathfrak{b}, \mathfrak{a})$, $0 \leq sim(\mathfrak{a}, \mathfrak{b}) \leq 1$, and $sim(\mathfrak{a}, \mathfrak{a}) = 1$.[1]

Rated similarity is a convenient means of defining *sim*. For example, subjects may be asked to position pairs of categories on a scale with endpoints "very dissimilar" and "virtually identical." A drawback to this procedure is that similarity ratings may implicitly recruit judgments of inductive inference; notably, two categories might be judged similar to the extent that they seem likely to share biological properties. In this case, including similarity as an input to our model of inductive inference threatens circularity. Instead of relying on ratings, we shall define $sim(\mathfrak{a}, \mathfrak{b})$ in terms of the overlap in brain activations produced by categorizing pictures of $\mathfrak{a}$ and $\mathfrak{b}$. Since neither "similarity" nor "probability" are mentioned during this procedure, the risk of circularity is minimized. The foregoing approach has already been applied to mammal categories (Weber and Osherson, to appear), and yields data consistent with the results reported here.[2]

We proceed as follows. The next section presents the function that maps unconditional probability and similarity to conditional probability. The probability data are then described, after which we turn to the experiment on neural similarity. The predictive success of the model is then considered, followed by discussion.

2 Theory

Our first step in constructing $Prob(\mathfrak{Q}\mathfrak{a} \mid \mathfrak{Q}\mathfrak{b}_1 \cdots \mathfrak{Q}\mathfrak{b}_n)$ is to adopt the following equation from the probability calculus.[3]

$$(3) \qquad Prob(\mathfrak{Q}\mathfrak{a} \mid \mathfrak{Q}\mathfrak{b}_1 \cdots \mathfrak{Q}\mathfrak{b}_n) = \frac{Prob(\mathfrak{Q}\mathfrak{a} \wedge \mathfrak{Q}\mathfrak{b}_1 \wedge \cdots \wedge \mathfrak{Q}\mathfrak{b}_n)}{Prob(\mathfrak{Q}\mathfrak{a} \wedge \mathfrak{Q}\mathfrak{b}_1 \wedge \cdots \wedge \mathfrak{Q}\mathfrak{b}_n) \quad + \quad Prob(\neg\mathfrak{Q}\mathfrak{a} \wedge \mathfrak{Q}\mathfrak{b}_1 \wedge \cdots \wedge \mathfrak{Q}\mathfrak{b}_n)}.$$

[1] The symmetry of similarity [$sim(\mathfrak{a}, \mathfrak{b}) = sim(\mathfrak{b}, \mathfrak{a})$] has been famously denied (Tversky, 1977 but see Aguilar and Medin, 1999). We adopt symmetry for simplicity in what follows.

[2] There is no overlap in participants between the two studies.

[3] For background in subjective probability, see Jeffrey (1983); Nilsson (1986); Halpern (2003). An elementary exposition is available at `http://www.princeton.edu/~osherson/primer.pdf`. In what follows, $\neg$ negates a statement, and $\wedge$ conjoins two statements.

To define $Prob(\mathtt{Q}\mathfrak{a} \mid \mathtt{Q}\mathfrak{b}_1 \cdots \mathtt{Q}\mathfrak{b}_n)$ it thus suffices to assign probabilities to the two conjunctions $\mathtt{Q}\mathfrak{a} \wedge \mathtt{Q}\mathfrak{b}_1 \wedge \cdots \wedge \mathtt{Q}\mathfrak{b}_n$ and $\neg\mathtt{Q}\mathfrak{a} \wedge \mathtt{Q}\mathfrak{b}_1 \wedge \cdots \wedge \mathtt{Q}\mathfrak{b}_n$). Let us begin with the former. Recall that we take the following quantities as given.

(4) (a) $Prob(\mathtt{Q}\mathfrak{a}), Prob(\mathtt{Q}\mathfrak{b}_1) \cdots Prob(\mathtt{Q}\mathfrak{b}_n)$

(b) $sim(\mathfrak{x}, \mathfrak{y})$, for $\mathfrak{x}, \mathfrak{y} \in \{\mathfrak{a}, \mathfrak{b}_1 \cdots \mathfrak{b}_n\}$.

Still relying on the probability calculus, the unconditional probabilities appearing in (4)a situate $Prob(\mathtt{Q}\mathfrak{a} \wedge \mathtt{Q}\mathfrak{b}_1 \wedge \cdots \wedge \mathtt{Q}\mathfrak{b}_n)$ between the following bounds, known to be the tightest possible (Neapolitan, 1990).

(5) $$\max\{0, Prob(\mathtt{Q}\mathfrak{a}) + Prob(\mathtt{Q}\mathfrak{b}_1) + \cdots + Prob(\mathtt{Q}\mathfrak{b}_n) - n\}$$

$$\leq \quad Prob(\mathtt{Q}\mathfrak{a} \wedge \mathtt{Q}\mathfrak{b}_1 \wedge \cdots \wedge \mathtt{Q}\mathfrak{b}_n) \quad \leq$$

$$\min\{Prob(\mathtt{Q}\mathfrak{a}), Prob(\mathtt{Q}\mathfrak{b}_1) \cdots Prob(\mathtt{Q}\mathfrak{b}_n)\}$$

In words, $Prob(\mathtt{Q}\mathfrak{a} \wedge \mathtt{Q}\mathfrak{b}_1 \wedge \cdots \wedge \mathtt{Q}\mathfrak{b}_n)$ is bounded above by the minimum of the unconditional probabilities $Prob(\mathtt{Q}\mathfrak{a}), Prob(\mathtt{Q}\mathfrak{b}_1) \cdots Prob(\mathtt{Q}\mathfrak{b}_n)$, and bounded below by the sum of these probabilities minus n (or zero if the latter subtraction yields a negative number). To illustrate, if $Prob(\mathtt{Q}\mathfrak{a}) = .6$, $Prob(\mathtt{Q}\mathfrak{b}_1) = .8$, and $Prob(\mathtt{Q}\mathfrak{b}_2) = .9$, then (5) implies: $.3 \leq Prob(\mathtt{Q}\mathfrak{a} \wedge \mathtt{Q}\mathfrak{b}_1 \wedge \mathtt{Q}\mathfrak{b}_2) \leq .6$.

We now rely on (4)b to choose a value for $Prob(\mathtt{Q}\mathfrak{a} \wedge \mathtt{Q}\mathfrak{b}_1 \wedge \cdots \wedge \mathtt{Q}\mathfrak{b}_n)$ within the interval (5). The key assumption is that the conjunction is more probable to the extent that its constituent categories are similar. The idea is that high similarity reduces the information expressed, e.g., it is less informative that peaches and nectarines share a given biological property than that peaches and avocados do. Specifically, letting p be the lower bound in (5), P the upper bound, and $S = \min\{sim(\mathfrak{x}, \mathfrak{y}) \mid \mathfrak{x}, \mathfrak{y} \in \{\mathfrak{a}, \mathfrak{b}_1 \cdots \mathfrak{b}_n\}\}$, we set

(6) $Prob(\mathtt{Q}\mathfrak{a} \wedge \mathtt{Q}\mathfrak{b}_1 \wedge \cdots \wedge \mathtt{Q}\mathfrak{b}_n) = [p \times (1 - S)] + [P \times S]$.

In words, the desired probability is the sum of the lower and upper bounds as weighted by the minimum similarity between categories in the conjunction. Continuing our example, if $sim(a, b_1) = .4$, $sim(a, b_2) = .6$, and $sim(b_1, b_2) = .8$ then $p = .3$, $P = .6$, $S = .4$, and $Prob(Qa \wedge Qb_1 \wedge Qb_2) = [.3 \times (1 - .4)] + [.6 \times .4] = .42$. Our use of minimum similarity will be justified shortly.

Consider now the second conjunction in (3), namely, $\neg Qa \wedge Qb_1 \wedge \cdots \wedge Qb_n$ with negated conjunct $\neg Qa$. To construct its probability on the basis of (4), we substitute $1 - Prob(Qa)$ for $Prob(Qa)$ in (4)a and $1 - sim(a, b_i)$ for $sim(a, b_i)$ in (4)b. Lower and upper bounds are then computed as before, along with minimum similarity. In our example, $Prob(Qa) = .6$, $Prob(Qb_1) = .8$, and $Prob(Qb_2) = .9$ so $Prob(\neg Qa) = .4$. The lower and upper bounds are thus $p = .1$ and $P = .4$, respectively, and $S = \min\{1 - sim(a, b_1), 1 - sim(a, b_2), sim(b_1, b_2)\} = \min\{.6, .4, .8\} = .4$. Hence, $Prob(\neg Qa \wedge Qb_1 \wedge Qb_2) = [.1 \times (1 - .4)] + [.4 \times .4] = .22$. Applying (3) yields:

$$Prob(Qa \mid Qb_1 \cdots Qb_n) = \frac{.42}{.42 + .22} = .65625.$$

Substituting $1 - sim(a, b_i)$ for $sim(a, b_i)$ makes sense in this situation inasmuch as the information in $Prob(\neg Qa \wedge Qb)$ varies inversely with the similarity of a and b (e.g., it is more surprising that peaches and nectarines do not share a given biological property than that peaches and avocados do not).

More generally, given a conjunction like $Qa \wedge \neg Qb_1 \wedge \neg Qb_2$ with one or more negated conjuncts, lower and upper bounds are computed from (4)a after substituting $1 - Prob(Qb_i)$ for $Prob(\neg Qb_i)$. Likewise, similarities are computed from (4)b by substituting $1 - sim(b_i, b_j)$ for $sim(b_i, b_j)$ if Qb_i and Qb_j appear in the conjunction with opposite polarities (i.e., just one of them negated). Illustrating with $Qa \wedge \neg Qb_1 \wedge \neg Qb_2$, the lower bound on its probability is

$$\max\{0, Prob(Qa) + (1 - Prob(Qb_1)) + (1 - Prob(Qb_2)) - 2\},$$

the upper bound is

$$\min\{Prob(Qa), 1 - Prob(Qb_1), 1 - Prob(Qb_2)\},$$

and the position of $Prob(Qa \wedge \neg Qb_1 \wedge \neg Qb_2)$ in this interval is determined by

$$\min\{1 - sim(a, b_1), 1 - sim(a, b_2), sim(b_1, b_2)\}.$$

Thus, the foregoing algorithm assigns probabilities to every conjunction for which the information in (4) is provided. Conditional probabilities may then be computed from the relevant ratio, as indicated by (3). To illustrate, the ratio for $Prob\,(\mathtt{Qa} \mid \mathtt{Qb})$ is

$$\frac{Prob\,(\mathtt{Qa} \wedge \mathtt{Qb})}{Prob\,(\mathtt{Qa} \wedge \mathtt{Qb}) \quad + \quad Prob\,(\neg\mathtt{Qa} \wedge \mathtt{Qb})}$$

whereas for $Prob\,(\mathtt{Qa} \mid \mathtt{Qb}_1, \neg\mathtt{Qb}_2)$ it is

$$\frac{Prob\,(\mathtt{Qa} \wedge \mathtt{Qb}_1 \wedge \neg\mathtt{Qb}_2)}{Prob\,(\mathtt{Qa} \wedge \mathtt{Qb}_1 \wedge \neg\mathtt{Qb}_2) \quad + \quad Prob\,(\neg\mathtt{Qa} \wedge \mathtt{Qb}_1 \wedge \neg\mathtt{Qb}_2)}.$$

In this way our model generates predictions for the queries in (1).

It is easy to verify that our model satisfies the following properties.

(7) (a) $Prob\,(\mathtt{Qa} \mid \mathtt{Qb}_1 \cdots \mathtt{Qb}_n)$ tends to increase with $Prob\,(\mathtt{Qa})$.

(b) $Prob\,(\mathtt{Qa} \mid \mathtt{Qb}_1, \mathtt{Qb}_2)$ tends to be greater for smaller values of $sim(\mathtt{b}_1, \mathtt{b}_2)$.

(c) $Prob\,(\mathtt{Qb}_1 \wedge \cdots \wedge \mathtt{Qb}_n) \geq Prob\,(\mathtt{Qb}_1 \wedge \cdots \wedge \mathtt{Qb}_n \wedge \mathtt{Qb}_{n+1})$.

(d) $Prob\,(\mathtt{Qb}_1 \wedge \cdots \wedge \mathtt{Qb}_i \wedge \cdots \wedge \neg\mathtt{Qb}_i \wedge \cdots \wedge \mathtt{Qb}_n) = 0$.

The second part of (7)a means that surprising (low probability) evidence has greater impact than obvious evidence. The first part affirms that prior probabilities influence conditional probability in the expected way. Property (7)b embodies the "diversity effect," discussed in Osherson et al. (1990); López (1995); López et al. (1997); Choi et al. (1997), and elsewhere. Property (7)c is the principle violated in studies of the "conjunction fallacy" (see Bonini et al., 2004; Tentori et al., 2004; Wedell and Moro, 2007 and references cited there). We do not expect such fallacies to arise for the simple stimuli at issue here. Finally, (7)d affirms that contradictory conjunctions are assigned zero probability. This last property is psychologically plausible so it is noteworthy that it fails to hold if the model's use of minimum similarity is replaced by average or maximum, for example. Likewise, (7)c depends on the use of minimum similarity.

3 Probability experiment

We collected estimates of conditional probability involving the following predicate and mammal categories.

PREDICATE: *are a significant source of at least* 10 *dietary vitamins*

CATEGORIES:				
	apples	avocados	bananas	grapes
	greenapples	strawberries	raspberries	tomatoes
	peaches	nectarines	pears	mangos
	lemons	limes	oranges	plums

Forty instances were constructed for each of the four types of queries illustrated in (1), respectively of forms $Prob(\mathtt{Qa} \mid \mathtt{Qb})$, $Prob(\mathtt{Qa} \mid \neg\mathtt{Qb})$, $Prob(\mathtt{Qa} \mid \mathtt{Qb}_1, \mathtt{Qb}_2)$, and $Prob(\mathtt{Qa} \mid \mathtt{Qb}_1, \neg\mathtt{Qb}_2)$. The specific categories figuring in a given instance were determined randomly under the constraint that that no two arguments have the same conclusion and premises. In addition, 16 unconditional probabilities were queried, one for each fruit [hence of form $Prob(\mathtt{Qa})$]. The 16 unconditional probabilities provide the information indicated in (4)a, needed to apply our theory.

The 160 conditional probability queries were randomly partitioned into four subsets of 40, each composed of ten queries of each type. Separate groups of eight Princeton undergraduates responded to the 40 queries of a given subset; each group also responded to the 16 unconditional queries (making 56 queries for each participant). Questions were presented via computer interface in random order. The concept of subjective probability (conditional and unconditional) was briefly explained prior to collecting data. All responses were constrained to fall in the unit interval (coded as percents). As illustration, the experiment included the four queries in (1).

The eight responses for each conditional probability were averaged, and likewise for the 32 responses for unconditional probabilities. Henceforth, these averages are called *the probabilities* (conditional and unconditional) of the corresponding queries. Statistics for the different kinds of items appear in Table 1. The probabilities for all 176 queries are available online via `http://www.princeton.edu/~osherson/supplement2.txt`.

4 Similarity

4.1 Overview

We now consider the $\binom{16}{2} = 120$ similarities needed as input to the model [see (4)b]. They were defined by comparing neural responses to a category-verification task. To

summarize our approach, let a given region of the brain be partitioned into voxels $v_1 \cdots v_m$. For fruit a and $i \leq m$, let $A(i, a)$ denote the metabolic activity in v_i provoked by the category-verification task involving a. We define the neural *distance* between categories a and b to be $\sum_i [A(i, a) - A(i, b)]^2$. Neural *similarity* arises from suitably inverting distance.

Our category-verification task was performed during fMRI scanning. We first present this paradigm then describe the brain regions chosen to define neural distance/similarity.

4.2 fMRI paradigm

Twenty-four chromatic images were collected for each fruit (half were mirror images of the other half). Corresponding to each of these "intact" images, we constructed a "scrambled" version resulting in a cloudy blur that preserves the brightness of the original.[4] Scrambled images served in the functional localizer described shortly.

The intact images figured in the following "match/mismatch" task. In each trial, the name of a fruit category was displayed for 2 seconds, followed by a sequence of fruit images. Each image was displayed for 333 ms with an inter-image interval of 333 ms. The sequence varied in length from trial to trial; in a given trial, there appeared $0 - 18$ images corresponding to the named category ("matching images"), followed by 3 images drawn from one of the remaining 15 categories, randomly chosen ("mismatch images"). In a given trial, participants were asked to press a key as soon as they noticed a mismatch image. Brain images collected between the onset of mismatch images and the end of a trial were not analyzed further (specifically, not modeled in the general linear model estimating activation in response to the images). Trials were separated by 10 seconds of fixation to allow the signal to return to baseline. By counterbalancing, each fruit received the same number of matching images. See the Appendix for further fMRI details.

Interleaved with the trials of the category-verification task were "visual control" trials involving the scrambled images. This task began with a 2 second presentation of the string #### followed by $6 - 12$ scrambled images, then 3 scrambled images each containing a low-contrast # sign. Participants were asked to press a key as soon

[4] All images can be retrieved from `http://www.princeton.edu/~osherson/fruit.tgz`.

as they saw an image with a # sign.

No fMRI participant performed the probability experiment, and any mention of similarity or probability was avoided until scanning was complete. After scanning, each participant used a computer interface to rate all 120 pairs of fruit categories for "conceptual similarity." For this purpose, they used a scale running from 0 ("dissimilar") to 100 ("similar"). Pairs were presented in individually random order.

Participants in the fMRI experiment were 12 Princeton University students and staff paid for their time (ages 18–32, 7 male).

4.3 Activation maps for each category

For every subject, an estimate of the neural response to pictures of a given category was generated in each voxel via multiple regression (details provided in the Appendix). For each category, the resulting activation maps were projected into Talairach space and averaged over the 12 subjects. The result was 16 activation maps, one for each fruit category. In the analyses subsequently described, subsets of these maps corresponding to various brain regions are used to calculate neural similarity.

4.4 Brain regions used to define similarity

From the activation maps for fruit categories, along with the activation map created from the visual control task, we identified voxels that were more responsive to intact fruit images than to their phase scrambled counterparts. Large clusters of such voxels were considered plausible regions for extracting neural similarities via squared differences, as described in Section 4.1.

Specifically, the contrast between a given voxel's response to intact versus scrambled images was measured by a 2×12 mixed ANOVA; the effect of intact vs. scrambled images was taken to be fixed, whereas differences among the 12 subjects were designated a random effect. Only voxels yielding $F(2, 12) > 19.75$ ($p < 0.001$), and more responsive to intact images, were retained. The resulting pool of voxels was then submitted to a clustering algorithm; voxels v_1 and v_2 were assigned to the same cluster just in case (a) v_1 and v_2 shared a face, or (b) v_1 shared a face with a voxel in a cluster containing

ν_2. Clusters smaller than 125 voxels were discarded. (Our method is standard; see, for example, Eger et al., 2008).

The foregoing procedure yielded 7 clusters ranging in volume from 127 to 5005 voxels. See Table 3. We concentrated on two large clusters that intersect Brodmann area 37 (BA37). One cluster is centered in the right middle occipital gyrus (Brodmann areas 19/37); the other is its counterpart in the left middle occipital gyrus. They occupy 5005 and 3140 voxels, respectively. Figure 1 indicates their positions.

The focus on BA37 stems from previous studies that document the role of the fusiform gyrus (the principal structure of BA37) in identifying category members and verifying their features (see, for example, Kan et al., 2003; Simmons et al., 2007). In Weber and Osherson (to appear), we describe an experiment that is isomorphic to the present one except that it involves 16 mammals categories rather than fruit. Left BA37 and regions close to the two clusters identified above emerged as particularly useful in defining neural similarity (as measured by success in predicting conditional probabilities). The two clusters, moreover, are close to regions known as left and right *lateral occipital complex* (LOC). These regions seem to play a central role in object identification (Grill-Spector et al., 2001). In sum, we concentrate on the four regions shown in Figure 1, henceforth identified as LBA37, RBA37, LLOC and RLOC. Other regions will be considered subsequently.

Let R be one of the brain regions chosen for study. Then R determines 120 neural distances, one for each pair of categories. Specifically, for categories a and b, its neural distance is the sum of $[A(i,a)-A(i,b)]^2$ where $A(i,a)$ is the activation of the i^{th} voxel of R in response to a (and likewise for b). To convert neural distances into similarities, we first normalized all distances to the unit interval, subtracted the distances from 1 to obtain similarities, then rescaled the similarities to the interval $[\frac{1}{3},\frac{2}{3}]$. The latter interval allows for the existence of fruit-pairs more similar than any figuring in our experiment, and other fruit-pairs that are less similar. The same interval figures in our earlier study (Weber and Osherson, to appear). It will be seen below that our results are robust over expansions of the interval.

For comparison with neural similarity, we also constructed *rated* similarity from the numerical similarities provided by the 12 subjects following fMRI scanning (see above). Each subject's 120 ratings were first normalized to the unit interval. For each of the

120 pairs, the 12 ratings were then averaged. Finally, the averages were normalized to the interval $[\frac{1}{3}, \frac{2}{3}]$, as for neural similarity.

Table 2 presents the Pearson correlations between the five sets of similarities (four neural, one behavioral). All four neural similarities correlate significantly but modestly with rated similarity ($p < .01$), the largest with RLOC ($r = .427$). The four neural similarities are highly correlated with each other. Figure 2 shows the distribution of the 120 similarities for the five kinds of similarity. We see that neural similarity is negatively skewed whereas rated similarity has positive skew.

5 Predicting conditional probability

Equipped with the two inputs listed in (4), our theory of inductive reasoning supplies a predicted response to all 160 queries in the probability experiment. One input — the 16 unconditional probabilities associated with each fruit category — is drawn directly from the probability experiment itself (see Section 3). The other input — the 120 similarities between pairs of fruit categories — is derived from the brain activity of any of our four regions (as discussed in Section 4.4), or directly from behavioral ratings.

5.1 Results for neural similarity: RLOC

When RLOC is used to define similarity, the predictions of our theory correlate at $r = 0.719$ ($p < .001$) with rated conditional probabilities; a scatterplot is provided in Figure 3. This correlation is robust to expansion of the similarity interval. Thus, changing the latter from $[\frac{1}{3}, \frac{2}{3}]$ to $[\frac{1}{4}, \frac{3}{4}]$ or to $[.15, .85]$ yields correlations of $r = 0.732$ and $r = 0.733$, respectively.[5] Contracting the interval to $[.4, .6]$ reduces the role of similarity in our theory, and lowers the correlation to $r = 0.679$. Similarity is effectively eliminated from the theory by use of the one-point interval $[.5, .5]$, fixing all similarities at one half. The probability of a given conjunction is then set to the midpoint of the interval determined by the unconditional probabilities of its conjuncts (see Section 2). Removing similarity from the theory in this way produces $r = 0.398$, which is reliably lower than the 0.719 that results from including similarity using $[\frac{1}{3}, \frac{2}{3}]$ ($p < .001$ via

[5] Changing the interval to $[.001, .999]$ yields $r = 0.727$ but the distribution of predicted probabilities is no longer smooth. Use of $[0, 1]$ produces division by zero.

a test for the difference between dependent correlations, Bruning and Kintz, 1977).[6] The latter contrast highlights the contribution of RLOC-based similarity in predicting rated probability.

The distribution of RLOC similarities is shown in Figure 2. The shape of the distribution, irrespective of the identities of particular fruit, might account for some or all of the predictive success of the theory. To determine whether category identity is a potent contributor to the correlation between predicted and rated conditional probabilities, we performed the following *permutation test*. For this test we randomly permuted the labels of the 16 fruit (e.g., renaming "apples" as "strawberries," "strawberries" as "nectarines," etc.), then applied our theory to the relabeled data. If category identity matters then the result of this procedure should be a lower correlation between predicted and rated probability. In 1000 permutations, only eight yielded a correlation as high as the one obtained without permutation (namely, $r = 0.719$, as reported above). We conclude that the mere distribution of similarities does not suffice to explain the predictive value of RLOC-based similarities. On the other hand, distribution is not irrelevant; for, the average correlation obtained under the 1000 permutations was .646 which is still superior to the $r = 0.398$ correlation that results from removing similarity from the theory altogether.

5.2 Results for other regions of interest

Among the regions of interest that we have identified (see Table 3), only LBA37, RBA37, RLOC and LLOC produce similarities that are significantly correlated with rated similarity. The performance of RLOC for predicting conditional probabilities was discussed above. LBA37 and RBA37 perform at essentially the same level ($r = 0.728, 0.718$, respectively) whereas LLOC performs somewhat less well ($r = 0.674$). Like RLOC, LBA37, RBA37 and LLOC produce correlations reliably higher than those emerging when fruit labels are randomly permuted. In 1000 trials, LBA37's original (unpermuted) correlation exceeded the permuted correlation 989 times (983 and 954 for RBA37 and LLOC, respectively). No other region of interest (Table 3) survives the permutation test even though all but one yield correlations that are significantly higher

[6] In what follows, all comparisons between correlations rely on this statistic.

than $r = 0.398$, obtained by removing similarity from the theory.[7]

The regions that predict rated similarity thus correspond to those that reliably predict conditional probability, namely, RLOC, LLOC, LBA37 and RBA37. It is thus surprising that rated similarity does not allow our model to predict conditional probability. In fact, the correlation between observed probability values and predictions based on rated similarity is only 0.333. Expanding the similarity interval from $[\frac{1}{3}, \frac{2}{3}]$ weakens the correlation further (e.g., $r = 0.268$ for $[\frac{1}{4}, \frac{3}{4}]$). Contracting the interval improves matters only slightly (e.g., $r = 0.410$ for $[.45, .55]$).

Results for other Brodmann areas

For completeness, we extracted similarity from every other Brodmann area in the brain (in addition to BA37). There was considerable variation in performance. Among the worst predictors were BA10 and BA35 (both left and right). Neural similarity from these areas was unrelated to rated similarity and produced correlations with probability judgments that did not survive the permutation test. (BA10 includes the frontal pole, and BA35 is perirhinal cortex.)

Left and right BA18, BA19 (occipital visual areas) perform best. Each is significantly correlated with rated similarity ($p < 0.01$), and survives the permutation test for conditional probability. (All four correlations with conditional probability exceed 0.70.) The right primary visual cortex RBA17 performs at the same level whereas its left homologue passes the permutation test but is less well correlated with rated similarity.

6 Discussion

We observed in the Introduction that rated similarity is a questionable input for theories of inductive inference (since it might embody tacit judgments of an inductive character). For this reason, we tested our theory of conditional probability using neurally based similarities elicited by a simple categorization task. Since no comparison of categories

[7]Observe that the outcome of the permutation test for a given region depends on more than the original, unpermuted correlation. Indeed, three regions listed in Table 3 fail the test despite having unpermuted correlations greater than LLOC, which passes.

was required in this task, neither inference nor similarity could have been on the minds of our fMRI subjects.

To identify neural regions of interest, we focussed on BA37 inasmuch as this area has been shown to participate in the representation of categories (see Section 4.4); also, the left side of this region emerges from our earlier study with mammal categories (Weber and Osherson, to appear). We also examined seven clusters of voxels that preferred intact to scrambled images, a natural criterion for narrowing the scope of our analysis. Two of those clusters, just posterior to BA37 and overlapping the left and right lateral occipital complex, strongly and reliably predict conditional probability. Regions of the parahippocampal and ventral fusiform cortices also preferred intact to scrambled images but did not predict these judgments better than permuted data. BA37 on both sides predicted inductive judgment with accuracy and reliability comparable to that of LOC. All four of these regions — namely, left and right LOC and BA37 — offer similarities that are well correlated with each other, and negatively skewed (see Figure 2). The shapes of the distributions play some role in their success at predicting probability. But the permutation tests reported above also show that the regions' sensitivity to category identity contributes to their performance.

The four regions that reliably predict conditional probability coincide with the regions that predict rated similarity (albeit modestly). It is therefore striking that rated similarity does not itself predict conditional probability, performing even worse than the theory with similarity removed. To explain this phenomenon we speculate that explicit similarity judgments are informed by subjects' theories about the internal structure of categories. For example, apples and peaches were rated as highly similar, possibly because both were recognized to be large and fleshy. Yet the two fruits might not be expected to share the property expressed by the predicate used in our experiment. Neural similarity may not be susceptible to such distortions.

As indicated at the end of the Results section, other visual areas (BA17, 18, 19) produce neural similarities that are significantly correlated with their rated counterparts. This raises a question about the kind of similarity underlying the ratings. We asked for "conceptual" similarity but it might have been predominantly visual (based on form and color). To explore the matter, we asked 15 additional subjects to rate all 120 pairs of fruits on "shape similarity" and separately on "color similarity." Conceptual

similarity was correlated with shape similarity at $r = 0.682$, and with color similarity at $r = 0.483$. Multiple regression of conceptual similarity on shape and color yields $r = 0.779$. These results confirm, unsurprisingly, that "conceptual" similarity of fruits depends partly on their shape and color. At the same time, shape and color perform no better in our theory than rated conceptual similarity. Using rated shape as the similarity source yields a correlation of just $r = 0.449$ with conditional probability; the correlation is $r = 0.141$ when color is used.

There remains the question of whether neural similarity is closer to rated conceptual similarity than to shape/color similarity. Of the nine regions that significantly predict conceptual similarity, only RLOC predicts shape similarity ($r = 0.33$, $p < 0.01$) or color similarity ($r = 0.20$, $p < 0.05$). All other correlations with shape/color ratings from these regions are less than $r = 0.13$. It thus appears that prediction of rated conceptual similarity is not mediated by the representation of shape or color.

In Weber and Osherson (to appear) we recovered neural similarities that predict rated similarities of mammals and, via our model, judgments of conditional probability. There is some overlap among the regions that predict successfully in the present and earlier studies, notably in left and right BA37, which pass the same permutation test in the experiment on mammals as they did in the fruit. Although LLOC predicted similarity and probability in mammals fairly well, neither it nor RLOC passed the permutation test in mammals. This may have been due to greater variability among the mammal images. Specifically, mammals were displayed in a variety of poses inducing diverse contours. In contrast, fruits are more simply structured, making their contours relatively homogeneous for a given category. Perhaps this greater homogeneity for fruit images allows a clearer representation of form by LOC. On the other hand, it is known that BA37 exhibits greater viewpoint invariance than more posterior structures (Vuilleumier et al., 2002), and represents semantic as well as perceptual aspects of categories (Kan et al., 2003; Wheatley et al., 2005; Simmons et al., 2007). BA37 might therefore be more robust to perceptual variability, explaining its predictive success in our two studies. In neither experiment could rated similarity be successfully exploited by our model of conditional probability.

Appendix: fMRI details

Image acquisition

Scanning was performed with a 3-Tesla Siemens Allegra fMRI scanner. Participants' anatomical data were acquired with an MPRAGE pulse sequence (176 sagittal slices) before functional scanning. Functional images were acquired using a T2-weighted echo-planar pulse sequence with 33 64 $\times$ 64-voxel slices, rotated back by 5 degrees on the left-right axis (axial-coronal $-5°$). Voxel size was $3 \times 3 \times 3$ mm, with a 1-mm gap between slices. The phase encoding direction was anterior-to-posterior. TR was two seconds (corresponding to the presentation of three images). Time to echo was 30 ms; flip angle was 90°. Field of view was 192×192 cm.

Stimuli

Sixteen fruit categories were presented in the course of the study, as indicated in Section 3. We collected color images illustrating each fruit, subtending about four degrees of visual angle. For each of the 16 target fruit, there were 24 such images (12 were mirror reversals of the other 12). Each functional run of the experiment also employed versions of the same fruit images in which the red, green, and blue layers of the images were separately phase-scrambled. (Phase scrambling preserves only the amplitudes of the Fourier spectrum of an image.) Some of the scrambled images were marked with a small, low-contrast pound (#) sign randomly placed in the image.

fMRI task

During scanning, participants performed several trials of an *experimental task* on intact stimuli and a *visual baseline* task on phase-scrambled stimuli. During a trial of the experimental task, participants saw the name of one of our 16 fruits for 2 s, then a series of 0–18 or fewer distinct, intact images of the named fruit, each presented for 333 ms with a 333-ms interstimulus interval (totaling up to 16 s of images). Each series was terminated by three distinct images of a fruit randomly drawn from one of the 15 remaining categories (thus, a mismatch to the initial label). Participants were instructed to press a key at the first appearance of an "intruder" image (that is, an image mismatched to the initial label). Thus, the participant might see nine apple images followed by three raspberry images, and be required to respond to the

first raspberry. Different trials displayed varying numbers of images before the intruder appeared, equated across trials for the 16 fruit. Only time points between the label and the first intruder were mapped to fruit regressors.

The images in the visual baseline task were phase-scrambled versions of the fruit pictures. Each baseline trial employed scrambled images from one fruit category. The form of the visual baseline task was identical to the main task, except that #### (in lieu of a category label) was presented prior to the images, and participants searched for a low-contrast crosshatch (#) in the sequence instead of a category mismatch. Images with # appeared at the end of each trial in positions corresponding to intruders in the main task. Only time points between the appearance of #### and the first image with # were mapped to visual baseline regressors. The study was organized into 8 runs, each including 12 experimental and 6 baseline trials.

All participants responded accurately on all trials.

Image analysis

Functional data were registered to the participant's anatomical MRI, despiked, smoothed with a 6 mm full-width at half-max Gaussian kernel, and normalized to percent signal change. For each participant, multiple regression was used to generate β values representing each voxel's activity in each fruit condition and each visual baseline condition. To calculate the β's, all variables were convolved with a canonical, double gamma hemodynamic response function and entered into a general linear model. Motion estimates were included as regressors of no interest. We relied on the statistical package AFNI (Cox, 1996) for preprocessing.

The 12 resulting activation maps for a given fruit (one for each subject) were projected into Talairach space and averaged, leaving us with 16 such maps, one for each fruit. Only voxels present in the intersection of all participants' intracranial masks were considered. (That is, when warped into Talairach space, different participants' brains may occupy slightly different volumes; we discarded voxels that were not occupied by everyone.) These average activation maps were the input to all subsequent analyses. Brodmann areas were identified by application of the "MRIcro" atlas (Rorden and Brett, 2000).

Tables

Table 1: Statistics on probabilities assigned to conditional and unconditional items

Type	*Number of items*	*Average*	*Standard Deviation*	*Minimum*	*Maximum*
$Prob\,(Qa \mid Qb)$	40	0.635	0.088	0.460	0.884
$Prob\,(Qa \mid \neg Qb)$	40	0.456	0.069	0.288	0.590
$Prob\,(Qa \mid Qb_1, Qb_2)$	40	0.673	0.065	0.551	0.840
$Prob\,(Qa \mid Qb_1, \neg Qb_2)$	40	0.521	0.104	0.249	0.744
$Prob\,(Qa)$	16	0.573	0.053	0.483	0.660
overall	176	0.572	0.115	0.249	0.884

Note: Each of the 160 conditional probabilities was evaluated by eight subjects. Each of the 16 unconditional probabilities was evaluated by 32 subjects.

Table 2: Intercorrelations between five sets of similarities

	RLOC	LLOC	LBA37	RBA37	Rated
RLOC	1.000	—	—	—	—
LLOC	0.718	1.000	—	—	—
LBA37	0.771	0.781	1.000	—	—
RBA37	0.765	0.727	0.936	1.000	—
Rated	0.427	0.282	0.262	0.265	1.000

Note: All correlations are significant ($p < .01$)

Table 3
Locations and prediction scores of clusters defined by functional localizer

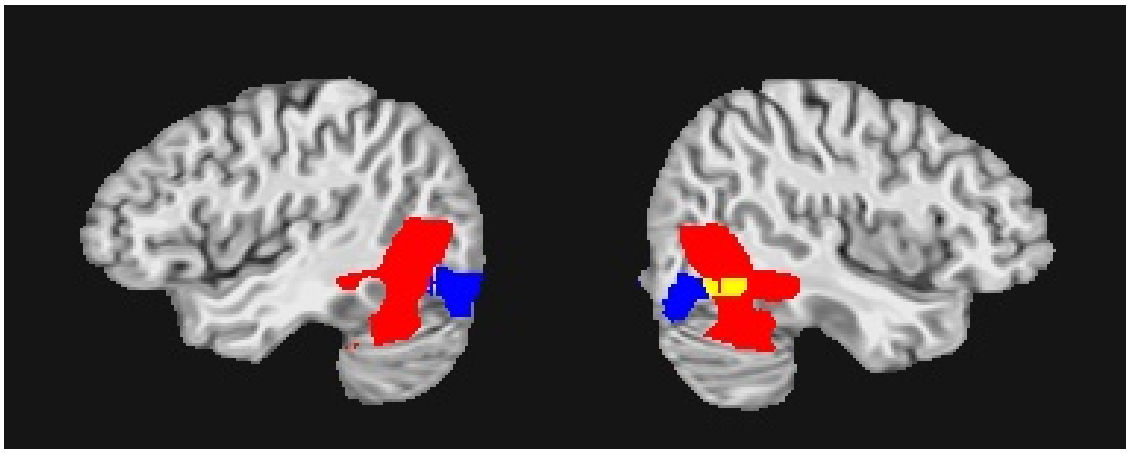
Gyrus	BA	Correlation with similarity	Correlation with probability	# voxels	x	y	z
Right middle occipital gyrus	19/37	**0.43**	**0.72**	5005	43	–71	–11
Left middle occipital gyrus	19/37	**0.28**	**0.67**	3140	–42	–76	13
Right BA 37	37	**0.26**	**0.72**	38,979	—	—	—
Left BA 37	37	**0.26**	**0.73**	38,614	—	—	—
Left lingual gyrus	17	0.19	0.69	185	–13	–93	–5
Left parahippocampal gyrus	36	0.12	0.70	489	–34	–30	–20
Right parahippocampal gyrus	36	0.12	0.68	1010	39	–35	–19
Left parahippocampal gyrus†	36	–0.08	0.42	127	–29	–18	–11
Left fusiform gyrus	37	–0.12	0.56	217	–47	–41	–22

†: Only this cluster fails to produce a correlation with conditional probability that is significantly better than that obtained when all similarities are set to 0.5.

Table 1: The clusters shown here were produced from the *intact* versus *scrambled* contrast (see the text). The x, y, z coordinates show centroids for each resulting cluster. Bolded correlations between rated and neural similarity are significant by t-test at $p < 0.01$. Correlations between rated probability and predicted probability are bolded in case they reliably exceed correlations based on permutation of labels (see the text). The probability that a given region fails this test is $p = 0.008$ for RLOC, $p = 0.046$ for LLOC, $p = 0.011$ for LBA37, and $p = 0.021$ for RBA37 (based on 1000 random permutations).

Figures

Figure 1: Neural sites of RLOC, LLOC, RBA37, and LBA37



Note: LOC is rendered in blue, BA37 in red; the area of overlap is yellow. The left image shows LLOC and LBA37 ($x = -42$ in Talairach space); the right shows their right-sided counterparts ($x = 42$).

Figure 2: Distributions of similarities

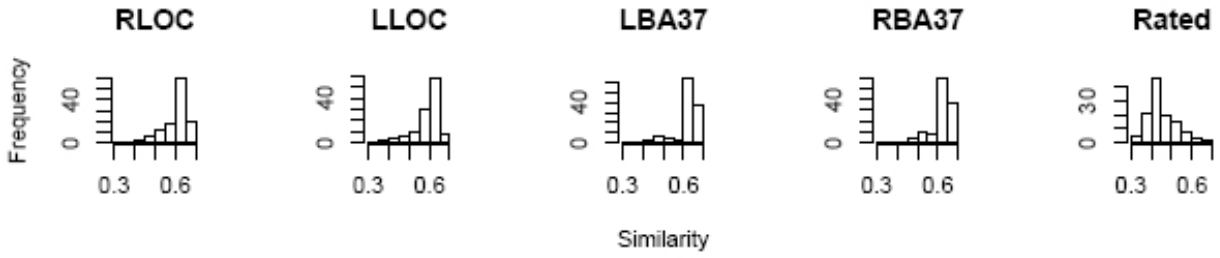
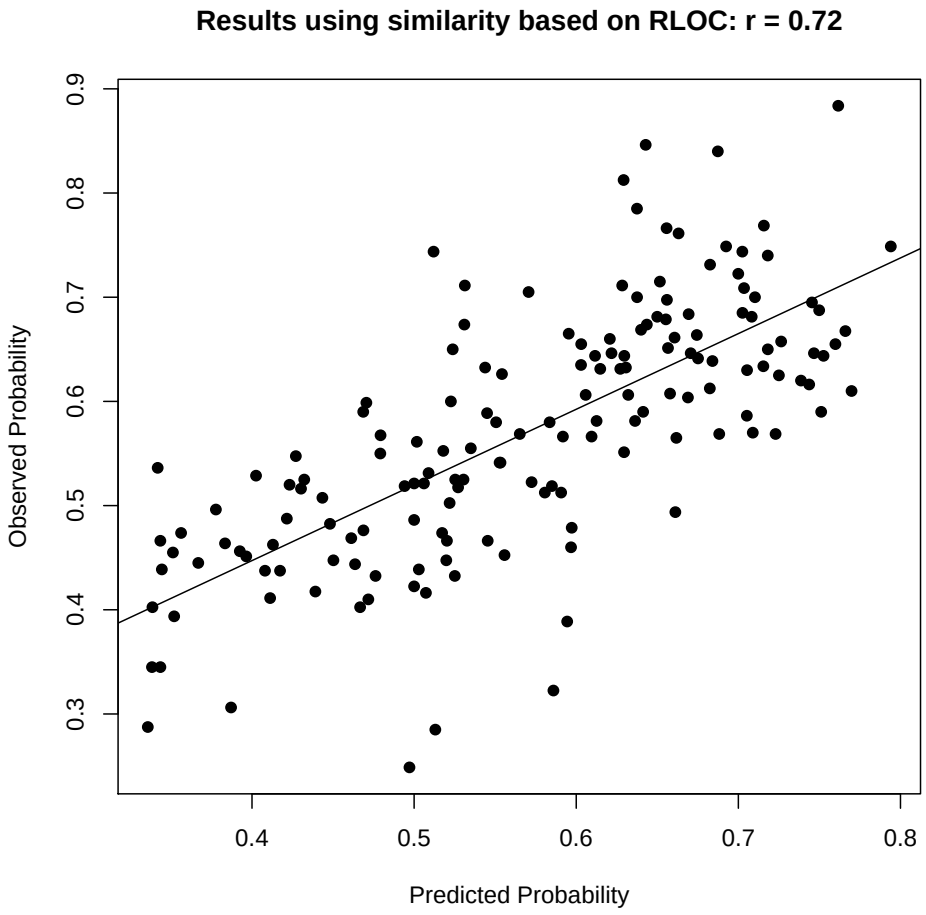


Figure 3: Observed versus predicted estimates of conditional probability based on RLOC



References

C. M. Aguilar and D. L. Medin. Asymmetries of comparison. *Psychonomic Bulletin & Review*, 6:328–337, 1999.

N. Bonini, K. Tentori, and D. Osherson. A different conjunction fallacy. *Mind and Language*, 19(2):199 210, 2004.

J. L. Bruning and B. L. Kintz. *Computational Handbook of Statistics*. Scott, Foresman and Co., Glenview IL, 2nd edition, 1977.

I. Choi, R. E. Nisbett, and E. E. Smith. Culture, Categorization and Inductive Reasoning. *Cognition*, 65:15 – 32, 1997.

R. W. Cox. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomed. Res.*, 29:162–73, 1996.

E. Eger, J. Ashburner, J.-D. Haynes, R. J. Dolan, and G. Rees. fMRI activity patterns in human LOC carry information about object exemplars within category. *Journal of Cognitive Neuroscience*, 20:356–370, 2008.

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

J. Y. Halpern. *Reasoning about Uncertainty*. MIT Press, Cambridge MA, 2003.

R. C. Jeffrey. *The Logic of Decision (2nd Edition)*. The University of Chicago Press, Chicago IL, 1983.

I. P. Kan, L. W. Barsalou, K. O. Solomon, J. K. Minor, and S. L. Thompson-Schill. Role of mental imagery in a property verification task: fmri evidence for perceptual representations of conceptual knowledge. *Cognitive Neuropsychology*, 20(3):525–540, 2003.

A. López. The diversity principle in the testing of arguments. *Memory & Cognition*, 23(3): 374 – 382, 1995.

A. López, S. Atran, J. Coley, D. Medin, and E. Smith. The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology*, 32: 251–295, 1997.

R. Neapolitan. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. John Wiley & Sons, New York NY, 1990.

N. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28(1):71–87, 1986.

D. Osherson, E. E. Smith, O. Wilkie, A. López, and E. Shafir. Category Based Induction. *Psychological Review*, 97(2):185–200, 1990.

C. Rorden and M. Brett. Stereotaxic display of brain lesions. *Behavioral Neurology*, 12: 191–200, 2000.

W. K. Simmons, V. Ramjee, M. S. Beauchamp, K. McRae, A. Martin, and L. W. Barsalou. A common neural substrate for perceiving and knowing about color. *Neuropsychologia*, 45: 2802–2810, 2007.

K. Tentori, N. Bonini, and D. Osherson. The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science*, 28:467 – 477, 2004.

K. Tentori, V. Crupi, N. Bonini, and D. Osherson. Comparison of confirmation measures. *Cognition*, 103(1):107–119, 2007.

A. Tversky. Features of Similarity. *Psychological Review*, 84:327–352, 1977.

P. Vuilleumier, R. N. Henson, J. Driver, and R. J. Dolan. Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. *Nature Neuroscience*, 5: 491–499, 2002.

M. Weber and D. Osherson. From similarity to inference. *Cognitive Science*, to appear.

D. H. Wedell and R. Moro. Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type. *Cognition*, 2007.

T. Wheatley, J. Weisberg, M. S. Beauchamp, and A. Martin. Automatic priming of semantically related words reduces activity in the fusiform gyrus. *Journal of Cognitive Neuroscience*, 17:1871–1885, 2005.