Embodied Verbal Semantics: Evidence from an Image-Verb Matching Task

Benjamin Bergen (bergen@hawaii.edu)
Department of Linguistics, 569 Moore Hall, 1890 East-West Rd.
Honolulu, HI 96822 USA

Shweta Narayan (shweta@icsi.berkeley.edu)
International Computer Science Institute, 1947 Center St., Suite 600
Berkeley, CA 94704-1198, USA

Jerome Feldman (jfeldman@icsi.berkeley.edu)
International Computer Science Institute, 1947 Center St., Suite 600
Berkeley, CA 94704-1198, USA

Abstract

It has recently been demonstrated that certain neural circuitry involved in the execution of specific motor actions is also used when the very same motor actions are observed or when language describing those actions is perceived. In humans, the pre-motor cortex is organized into regions that are involved in the execution and observation of actions performed by at least the following three general areas: the mouth, the hand, and the leg. The discovery of this "mirror system", involved in production and perception of motor behavior, leads to a viable hypothesis about the processing of linguistic units that refer to these actions. It could be that understanding a verb describing an action involves the activation of the very same mirror circuitry involved in performing and recognizing that action. This hypothesis is tested in a matching task, in which subjects were presented first with an image depicting some action, followed by a verb that either described that action or did not. They were asked to decide as quickly as possible whether the verb was appropriate to the image. It was reasoned that if the verbs and images for particular actions recruited the same mirror circuitry, then there should be interference in those cases where the actions described by the verb and image were not the same but used the same effector. The results showed that it took subjects significantly longer to reject non-matching verbs and images when the two shared an effector than when they did not. These results support the hypothesis that understanding action language requires the activation of effector-specific neural circuitry in the human mirror system.

Introduction

An important recent development in our understanding of how actions are perceived is the discovery of so-called "mirror neurons" (Gallese et al 1996, Rizzolatti et al. 1996). Mirror neurons are cells in the monkey cortex that are selectively activated during the performance of specific motor functions, but which also become active when the individual perceives another person or monkey performing the same function. There are no single unit studies in humans, but comparable "mirror activity" patterns have been demonstrated with imaging studies. Mirror circuits have thus been shown to serve dual roles in producing actions and recognizing these actions performed by others.

It has also been established that this mirror system extends to the somatotopic organization of the pre-motor and parietal cortex (Buccino et al 2001). In particular, the execution or observation of actions produced by the mouth, leg, and hand activate distinct parts of pre-motor cortex, found in ventral sites, dorsal sites, and intermediate foci, respectively. When appropriate target objects are present, there is also activation in a somatotopic activity map in parietal cortex.

These results bear upon a key question in the study of language - how motion verbs are processed by language users. Does understanding a motion verb entail any detailed internal simulation of motion? Are the areas of the brain responsible for enacting motor actions activated for this purpose?

Two recent studies provide evidence that processing motion language associated with particular body parts results in the activation of areas of motor and pre-motor cortex involved in producing motor actions associated with those same effectors. Using both behavioral and neurophysiological evidence, Pulvermüller et al. (2001) found that verbs associated with different effectors were processed at different rates and in different regions of motor cortex. In particular, their results showed that subjects performing a lexical decision task respond to verbs referring to mouth actions faster than they do to verbs involving the leg. The researchers also found that the areas of motor cortex involved in leg and mouth motion received more activation during the processing of leg- and mouth-related words, respectively. More recently, Tettamanti et al. (m.s.) have shown through an imaging study that passive listening to sentences describing mouth versus leg versus hand

motions activate different parts of pre-motor cortex (as well as other areas, specifically BA 6, BA 40, and BA 44).

Both of these studies confirm that motor representations specific to particular effectors become active when subjects are exposed to linguistic input. But there are several additional issues that should be investigated. First, it remains to be determined to what extent the activation of these neural representations is recruited for the purpose of extracting meaning from linguistic utterances – that is, for constructing meaning. It could be the case, for example, that the activation and their effects found by Tettamanti et al. and Pulvermüller et al. are simply associated, collateral patterns of activation, which play no functional role in the interpretation of language.

A second remaining research question is whether those motor structures that are activated during the processing of language pertaining to motor action are the same structures that have been demonstrated to become active during the visual perception of motor actions. The evidence is mounting that those areas of motor and premotor cortex that are specialized for particular motor actions also become active during the processing of visual and linguistic inputs corresponding to those actions. But it has not yet been determined whether those two modes of input are at any point processed by the same circuitry.

The experimental study reported in this paper addressed these two issues by investigating whether there is interference between visual and linguistic input during the process of matching images and verbs that depict related actions.

The experiment used a matching paradigm, in which subjects were first presented for one second with a stick-figure image and then were asked to decide as quickly as possible if a verb they subsequently saw on the screen was a good description of that image or not. The images and verbs all depicted actions that were primarily associated with one body area – the hand, the mouth, or the leg. In half of the trials, the verb was a good description of the image (the 'matching' condition) and in the other half of the trials, it was not (the 'non-matching' conditions).

If the process of understanding motion verbs makes use of the same neural resources that are used for recognizing motion itself, then matching a verb with an image should require a single coherent neural representation to win out over competitors. When a verb and image do not match, the neural structures involved in the recognition of each should become active. In general, in order for a neural system to function smoothly, there must be inhibition between structures responsible for similar but incompatible functions, and functionally related neural structures with more similar representations should mutually inhibit each other more strongly than less closely related neural structures. We can therefore hypothesize that there should be more interference in identifying a mismatched verb and image when they use the same effector than when an image and verb that use different effectors are compared.

In terms of the task at hand, the subjects should take more time to reject non-matching verbs when those verbs make use of the same major body part as the image than when they made use of different body parts. Thus, we separated the non-matching pairs into two sets, those that had the same body part (non-matching same effector) and those that had different body parts (non-matching different effector). To clarify, examples of stimuli in each condition are shown in Figure 1 and 2 below.

Figure1: Verbs in the three conditions for the image *run*

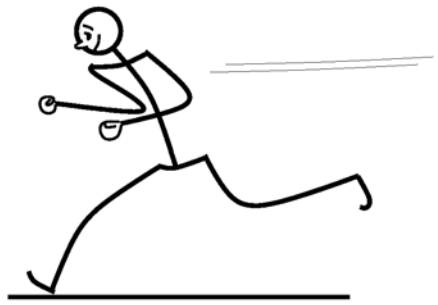
	 *Image*
Matching verb	'run'
Non-matching, same effector	'kick'
Non-matching, diff. effector	'drink'

Figure 2: Verbs in the three conditions for the image *scratch*

	 *Image*
Matching verb	'scratch'
Non-matching, same effector	'hold'
Non-matching, diff. effector	'stumble'

Method

Each trial consisted of a visual stimulus like the images shown above, which was presented for one second, followed immediately by a 500 millisecond interstimulus interval, the first 450 milliseconds of which included a visual mask covering the whole screen. This was meant to reduce any

priming effects that resulted from visual imagery. An English verb in written form was then presented until the subject pressed a button indicating that the verb was or was not a good description for the action depicted in the image. The verb fell into one of the three conditions described above: (1) matching, (2) non-matching same effector, and (3) non-matching different effector.

Subjects were 39 members of the University of California at Berkeley community, all native speakers with normal hearing and language competence and with normal or corrected-to-normal vision. They received course credit in exchange for participation in the experiment. Subjects were given the following instructions:

> This experiment tests how people relate words and images. You will first see an image of a person performing an action. Then you will see a verb that is either a good description of the action or not. Your job is to decide as quickly as possible whether or not the word is a good description of the action. If the word is a good description of the action, press the 'Yes' button. If it is not, press the 'No' button.

After a brief practice session that included a total of 14 image-word trials, each of the 39 subjects was randomly placed in one of two groups. Each subject was presented with each image a total of two times, once in each of two halves of the experiment. Each saw (1) each image followed by a matching verb once, (2) half of the images with a non-matching different effector verb and (3) the other half of the images with a non-matching same effector verb. The verbs were distributed such that each image that was shown to one group in the non-matching same effector condition was in the non-matching different effector condition for the other group.

A total of 16 stick-figure images representing each of the three effectors – mouth, hand, and leg – were hand-created by one of the experimenters using a graphics editor. Each of these images was intended to specifically depict a particular type of motor action. Aside from posture and occasional movement lines, head and eye position, as well as overall body shape had to be manipulated to evoke actions that were as specific as trip and scream. Many of the actions depicted by these images thus also involved some movement of adjacent body parts.

Verbs that appropriately described these images were selected using a pre-test, in which 13 subjects, all native speakers of American English, were presented with each image, and were asked to provide the verb they thought best described the action depicted by the image. The most frequent response to each image was taken as the matching verb. Given the 48 images and their matching verbs, each image was then randomly assigned one of the verbs that matched another verb with the same effector and one of the verbs with another effector. These were the non-matching same effector and non-matching different effector verbs for that image, respectively. Each verb was used twice in the matching condition (once for each subject group) and once in each of the non-matching conditions, so these three conditions were completely balanced.

Results

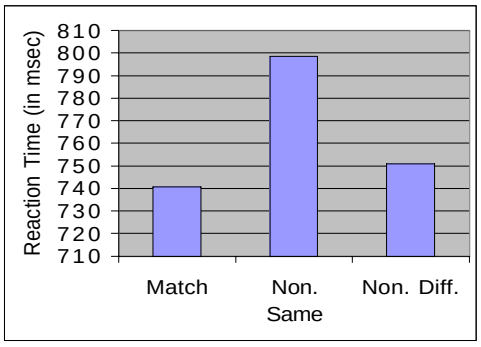
There fewer than 2% incorrect responses overall, and there were no significant differences in errors among the conditions. In what follows, we consider only those responses that were correct – that is, only 'yes' responses to the matching condition and only 'no' responses to the two non-matching conditions. In order to ensure that any significant differences did not result from a small set of outliers, we also removed all reactions that deviated more than 2 standard deviations from the mean for that trial. Of those data that remained, the means were different in the three conditions, as shown in Table 1 below, where we can see that the mean reaction time to non-matching verbs is on average 48 milliseconds longer when the verb and image use the same effector than when they use different effectors.

Table 1: Means Table for Reaction Time as product of condition

	Count	Mean (msec)	Std. Dev.	Std. Err.
Match	1596	740.57	257.54	6.45
Non-matching Same Effector	840	798.54	251.00	8.66
Non-matching Diff. Effector	870	750.93	204.74	6.94

This can also be seen graphically in Figure 3 below.

Figure 3: Reaction Time by condition



The difference between the conditions as well as subject identity were statistically significant in an ANOVA, as seen in Table 2 below.

Table 2: ANOVA Table for Reaction Time

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
SUBJ	38	1940068.274	51054.428	17.033	<.0001	647.253	1.000
COND	2	76079.476	38039.738	12.691	<.0001	25.382	.998
Residual	76	227801.487	2997.388				

The direction of the significant difference between the two non-matching conditions conforms to the hypothesis that the rejection of a verb as an image description is delayed when the same effector is involved in both.

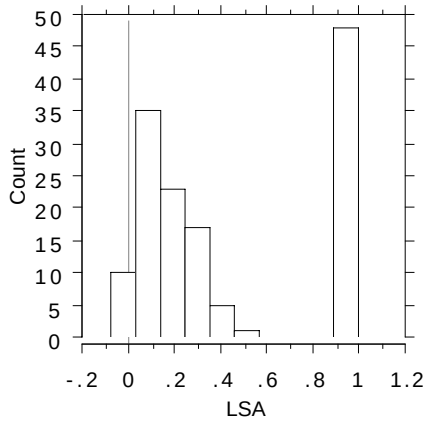
An alternative explanation for this behavior would be that it is semantic similarity in a general sense rather than the sharing of an effector that gives rise to this effect. Since actions that share an effector are in general similar to each other in dimensions other than the identity of the effector, it might be that subjects simply took longer to reject verbs that described actions that were in some way more similar to the action depicted by the images the followed.

We addressed this concern by evaluating the effect of semantic similarity of the presented verb with the verb that was most commonly associated with the particular image in the pretest described above. In other words, for the three examples in Figure 1, we determined a semantic similarity score between run and run (matching), between run and kick (non-matching, same effector) and between run and drink (non-matching, different effector). This is an indirect way of evaluating the similarity between an image and a verb, since it is mediated by a verb describing the image, but for the time being it may have to do in the absence of more direct methodologies.

The semantic similarity metric we used was a similarity rating produced by Latent Semantic Analysis (LSA - Landauer et al. 1998, and http://lsa.colorado.edu/). LSA, among other things, is a statistical method for extracting and representing the similarity between words or texts on the basis of the contexts they do and do not appear in. Two words or texts will be rated as more similar the more alike their distributions are. LSA has been shown to perform quite like humans in a range of behaviors, including synonym and multiple-choice tasks. Of relevance to the current discussion is the pairwise comparison function, which produces a similarity rating from -1 to 1 for any pair of texts. Identical texts have a rating of 1, while completely dissimilar ones would have a rating of -1.

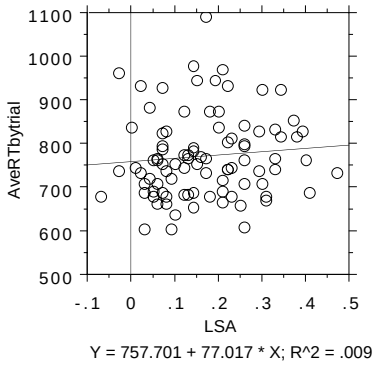
With LSA ratings assigned to each trial, we took the average RT per trial (that is, per image-verb pair) and performed a regression analysis with the LSA rating for the verb and picture's most plausible verb, as described above. This regression included only the non-matching conditions, as including the matching condition (with LSA ratings of 1) produces an abnormal distribution, as seen in Figure 4.

Figure 4: Histogram of LSA ratings for all word pairs



Considering only the two non-matching conditions, there was a very weak correlation between LSA rating and reaction time (R = 0.094). As seen from the regression graph in Figure 5 below (and from the positive value of the regression coefficient R), the relation is in the predicted direction - the trend is for subjects to take longer to reject more similar pairs of words and pictures than less similar ones.

Figure 5: Regression of RT per trial by LSA



However, this trend is insignificant, with p=.378. So while the similarity between a non-matching verb and image as measured by LSA qualitatively seems to account for a small amount of the variance in reaction time, it does not do so significantly. Of course, this does not prove that sharing an effector and not other sorts of similarity is responsible for the reaction time effects we've seen. The LSA rating might be a flawed measure of similarity in general or with respect to verb-image similarity. For this reason, further studies like the ones described below will be required to test whether the effects are actually based on effector identity. The absence of a significant relation between LSA rating and reaction time shown by the regression above does, however, suggest that overall similarity does not transparently account for the interference behavior we found.

Discussion

Subjects took significantly longer to reject a verb that did not describe an action depicted by an image when that verb shared an effector with the image it followed. This result provides evidence that, when understanding motion verbs, language users recruit some resources normally used for perceiving motion in general. This finding also supports the idea that the understanding of motion verbs depends on the active simulation or imagination of motion. Finally, it is also congruent with findings that motor actions are perceived in part through activating some brain circuits involved in motor control of those actions and that the comprehension of verbs denoting motor actions also employ those same motor circuits.

The major interference effect between images and verbs that describe different actions performed by the same effector could arise from interaction within the neural

structures in question. Specifically, it could be that the mirror circuits involved in the recognition of verbs and images are specific to particular types of action. This specificity has been seen in monkeys (Gallese et al. 1996), where a given mirror neuron may code a specific type of gesture, like a precision grip, for example. If this is the case, then the more similar two actions are, the more necessary it will be for mirror circuits that encode those actions to mutually inhibit each other. This mutual inhibition may give rise to the sort of delay in rejection we saw evidenced in the experiment described above in the following way.

We know from Pulvermüller et al. (2001) and Tettamanti et al. (m.s.) that words pertaining to motor actions yield activation of the specific mirror structures that are responsible for performing those motor actions. Words and images most likely also provoke activation of closely related mirror structures, because of their similar perceptual character - crucially sharing effectors, but also other features like weight distribution, overall body position, and so on. Neural representations for actions using the same effector will thus be co-activated whenever a linguistic or pictorial input of a motor action is presented. Because neural structures representing closely similar actions must mutually inhibit each other, a single mirror structure eventually wins out over others, leading to the perception of the action. In the matching task described in this paper subjects were asked to take two inputs and determine whether they were the same. A crucial part of this process must be for the subject to determine if they have perceived one action or two. At the neural level, this translates into the strong activation of one (in the matching condition) versus two mirror structures (in the non-matching conditions). When in the non-matching condition, the two mirror structures are very similar, for example, when they share an effector, they will strongly inhibit each other, and it will thus take longer in such a trial for two distinct active mirror structures to emerge, and therefore for a subject to arrive at two distinct action perceptions. By comparison, when the actions are unlike each other, for example jog and laugh, there will be less mutual inhibition, and the two mirror structures will take less time to become co-active.

This proposed mechanism yields hypotheses that can be tested in future work. First, the mutual inhibition of related structures being invoked to explain the delay in rejection of matching should extend to other tasks as well. That is, if it truly is the case that an active representation for a given action is slowing down the activation of another representation through inhibition, then activation of an image should also delay identification or categorization of a word or image depicting a closely related action that is presented simultaneously or with a very short inter-stimulus interval. In other words, this delay in matching should be a generalized priming effect, as long, that is, as the prime can be ensured to remain active during the recognition of the target.

Second, if the mutual inhibition mechanism is truly motor-representation-specific and does not result from the particular modality in which stimuli are presented, then reversing the order of prime and target should yield the same result as described above. Subjects could be presented first with a verb and then with an image, and would be asked to perform the same task as the subjects in the current experiment. Such an experiment is currently being planned.

Regardless of the exact neural mechanisms responsible for them, the results reported here may illuminate a key question in the study of language: what does it mean for a language user to understand a linguistic form or utterance? It has been hypothesized in various places (cf. Feldman et. al 1996, Narayanan 1997, and Bailey et al. 1998 as predecessors of MacWhinney 1999 and Bergen & Chang In Press,) that deep language understanding results from the enactment of an internal simulation of the content of the utterance. For example, in order to understand an utterance like “John threw the water balloon”, language users might be activating some subset of the motor structures responsible for that particular type of throwing. Some evidence for this hypothesis comes from the remarkable ability language users have to make immediate inferences about actions they have heard described. For example, upon hearing “John threw the water balloon”, the understander can immediately answer questions about the hand shape used, the amount of pressure applied to the balloon, the trajectory of the arm, and so on.

The evidence is mounting that recalling language associated with particular perceptual or motor functions activates the neural areas involved in those same functions (Pulvermueller et al. 2001, Tettamanti et al. M.s.). If this is the case, linguistic activation of these areas is just one among a number of different uses they serve. Recalling actions and also motor imagery have also been shown to result in activation of motor circuitry.

Recent work by Nyberg et al. (2001) provides compelling evidence that recalling actions activates the same brain areas as encoding them does. Nyberg et al. presented subjects with a verbal command, like look at your hand, and asked them to execute or imagine executing the motor action. This yielded activity, among other places, in motor and parietal cortex. When the subjects were later asked to provide the direct object of the verbs they had either heard and enacted or imagined enacting, similar brain regions, including motor and parietal cortex, showed differential activation.

It is not only recalling actions that yields activation of motor areas of the brain – motor imagery does, as well. A number of studies over the past twenty years have demonstrated through a variety of methods that the brain areas concerned with motor control are also activated during motor imagery (Roland et al 1980, Porro et al. 1996). For example, Lotze et al. (1999) recently found in an fMRI study that motor areas, including supplementary motor cortex, premotor cortex, and motor cortex were all activated in both motion and in imagined motion. Other areas involved in both executed and imagined motion are the cerebellum (Decety et al. 1990) and parietal cortex (Sigiru et al 1996).

Together, the demonstrated use of motor structures for imagined action, recalled action, and now for language processing lend credence to a view of meaning and thought that is tightly grounded in the experiences a person has interacting with the world around them. The findings reported here support an embodied theory of the meanings of linguistic units and the utterances they appear in – one in which motor language has meaning through reference to experiences that the individual can evoke. An extension of this theory suggests that abstract words derive their meanings from metaphorical and other projections to these same embodied experiences and we are planning additional experiments to test the extension.

The findings reported here, which suggest that language activates motor circuitry involved in producing and perceiving actions, provide evidence for the embodiment of linguistic meaning and the important role of the particularities of human neural circuitry to theories of the language understanding process.

Acknowledgments

We would like to thank Srini Narayanan, Teenie Matlock, Lokendra Shastri, Carter Wendelken, and other members of the Neural Theory of Language group, as well as Art Glenberg, Brian MacWhinney, and three anonymous reviewers for their helpful comments and suggestions on this work.

References

Bailey, D., Chang, N., Feldman, J., & Narayanan, S. (1998). Extending Embodied Lexical Development. Proceedings of the Twentieth *Annual Meeting of the Cognitive Science Society COGSCI-98*, Madison.

Bergen, B. & Chang, N. (To Appear). Embodied Construction Grammar in Simulation-Based Language Understanding. In J-O. Östman and M. Fried (Eds.), *Construction Grammar(s): Cognitive and Cross-language dimensions*. John Benjamins.

Buccino, G, Binkofski, F, Fink, G.R., Fadiga, L, Fogassi, L, Gallese, V, Seitz, R.J., Zilles, K, Rizzolatti, G, & Freund, H.J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *European Journal of Neuroscience 13(2)*: 400-404.

Decety J, Sjoholm, H, Ryding, E, Stenberg, G, & Ingvar, D.H. (1990). The cerebellum participates in mental activity: tomographic measurements of regional cerebral blood flow. *Brain Res 535:*313–317.

Feldman, J., Lakoff, G., Bailey, D., Narayanan, S., Regier, T., & Stolcke, A. (1996). L0: The First Five Years. *Artificial Intelligence Review*, *v10* 103-129.

Gallese , V., Fadiga, L., Fogassi, L., & Rizzolatti, G.. 1996. Action recognition in the premotor cortex. *Brain 119*: 593-609.

Landauer, T., Foltz, P., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259-284.

Lotze, M., Montoya, P., Erb, M., Hülsmann, E., Flor, H., Klose, U., Birbaumer, N., & Grodd, W. (1999) Activation of cortical and cerebellar motor areas during executed and imagined hand movements: An fMRI study, *Journal of Cognitive Neuroscience, 11(5)*: 491-501

MacWhinney, B. (1999). The emergence of language from embodiment. In B. MacWhinney (Ed.), *The emergence of language*. Mahwah, NJ: Lawrence Erlbaum.

Narayanan, S. (1997). Talking The Talk Is Like Walking the Walk: A Computational Model of Verbal Aspect. *Proceedings of the Nineteenth Annual Meeting of the Cognitive Science Society COGSCI-97.* Stanford: Stanford University Press.

Nyberg, L., Habib, R., McIntosh, A. R., & Tulving, E. (2000). Reactivation of encoding-related brain activity during memory retrieval. *Proc. Natl. Acad. Sci. USA 97:* 11120–11124.

Nyberg, L., Petersson, K.-M., Nilsson, L.-G., Sandblom, J., Åberg, C., & Ingvar, M. (2001). Reactivation of motor brain areas during explicit memory for actions. *NeuroImage, 14,* 521-528.

Porro CA, Francescato MP, Cettolo V, Diamond ME, Baraldi P, Zuian C, Bazzocchi M, & di Prampero PE (1996) Primary motor and sensory cortex activation during motor performance and motor imagery: a functional magnetic resonance imaging study. *J Neurosci 16*:7688–7698.

Pulvermueller, F., Haerle, M., & Hummel, F. (2001). Walking or Talking?: Behavioral and Neurophysiological Correlates of Action Verb Processing *Brain and Language 78*, 143–168.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L.. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*:131--141.

Roland PE, Larsen B, Lassen NA, & Skinhoj E (1980) Supplementary motor area and other cortical areas in organization of voluntary movements in man. *J Neurophysiol 43*:118–136.

Sirigu A, Duhamel JR, Cohen L, Pillon B, Dubois B, Agid Y (1996). The mental representation of hand movements after parietal cortex damage. *Science 273:*1564–1568.

Tettamanti, M., Buccino, G., Saccuman, M.C., Gallese, V., Danna, M., Perani, D., Cappa, S.F., Fazio, F., & Rizzolatti, G. (Unpublished Ms.) Sentences describing actions activate visuomotor execution and observation systems.

Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proc. Natl. Acad. Sci. USA 97*: 11125–11129.