NeuroImage 152 (2017) 619–627

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

An asymmetrical relationship between verbal and visual thinking: Converging evidence from behavior and fMRI

Elinor Amit[a,b,*], Caitlyn Hoeflin[c], Nada Hamzah[a], Evelina Fedorenko[b,d,**]

[a] Brown University, United States
[b] Massachusetts General Hospital, United States
[c] Massachusetts Institute of Technology, United States
[d] Harvard Medical School, United States

ARTICLE INFO

Keywords:
Inner speech
Visual imagery
Modes of thought
fMRI

ABSTRACT

Humans rely on at least two modes of thought: verbal (inner speech) and visual (imagery). Are these modes independent, or does engaging in one entail engaging in the other* To address this question, we performed a behavioral and an fMRI study. In the behavioral experiment, participants received a prompt and were asked to either silently generate a sentence or create a visual image in their mind. They were then asked to judge the vividness of the resulting representation, and of the potentially accompanying representation in the other format. In the fMRI experiment, participants had to recall sentences or images (that they were familiarized with prior to the scanning session) given prompts, or read sentences and view images, in the control, perceptual, condition. An asymmetry was observed between inner speech and visual imagery. In particular, inner speech was engaged to a greater extent during verbal than visual thought, but visual imagery was engaged to a similar extent during both modes of thought. Thus, it appears that people generate more robust verbal representations during deliberate inner speech compared to when their intent is to visualize. However, they generate visual images regardless of whether their intent is to visualize or to think verbally. One possible interpretation of these results is that visual thinking is somehow primary, given the relatively late emergence of verbal abilities during human development and in the evolution of our species.

Introduction

What is the nature of the representations that mediate human thought* Two representational formats are most commonly discussed in the literature (e.g., Baddeley and Hitch, 1974; Paivio 1986). First, people talk to themselves silently – a phenomenon often termed *inner speech* (Vygotsky, 1962, 2012; Zivin, 1979; Sokolov, 1972). Inner speech has been shown to play an important role in propositional thought (e.g., Carruthers, 2002; Pléh, 2002), working memory (e.g., Baddeley and Hitch, 1974), long-term memory (e.g., Schrauf, 2002), numerical cognition (e.g., Frank et al., 2012), self-awareness (e.g., Siegrist, 1995), and self-reflection (e.g., Morin and Michaud, 2007). Although some have discussed inner speech in a narrow sense of phonological-level processes (e.g., Price, 2012; Smith et al., 1998), we adopt a broader definition, which includes generation of meaningful linguistic representations (e.g., Delamillieure et al., 2010; Vygotsky, 2012). The second type of representation is *visual imagery*. Visual imagery occurs when perceptual information is accessed from memory, giving rise to the experience of "seeing with the mind's eye" (e.g., Ganis et al., 2004), and has been shown to be important in simulating object manipulation (e.g., Shepard and Metzler, 1971), episodic memory (e.g., Paivio, 1986), and self-projection (e.g., Buckner and Carroll, 2007).

Given that humans rely on both the verbal and the visual modes of thought (e.g., Amit et al., 2009, 2013; Paivio, 1986), a question naturally arises about the relationship between the two. In this paper we ask whether these two modes of thought are independent of each other. If so, people should be able to engage in visual imagery without engaging in inner speech, and vice versa. Alternatively, it may not be possible to engage in one mode of thought without invoking the corresponding representation in the other. Yet another possibility is that an asymmetry exists between inner speech and visual imagery. It may be impossible to verbalize without invoking a corresponding visual representation (e.g., Barsalou, 2010; Boroditsky and Prinz, 2008; Hume, 1739/1951). Or it may be impossible to create a visual image without an accompanying "voice over" (e.g., Dennett, 1991; Paivio, 1986). The former may be a priori more likely given that visual imagery

* Corresponding author at: Brown University, United States.
** Corresponding author at: Massachusetts General Hospital, United States.
E-mail addresses: elinor.amit@gmail.com (E. Amit), evelina9@mit.edu (E. Fedorenko).

http://dx.doi.org/10.1016/j.neuroimage.2017.03.029
Received 19 September 2016; Accepted 15 March 2017
Available online 18 March 2017
1053-8119/ © 2017 Elsevier Inc. All rights reserved.

has emerged earlier evolutionarily, and precedes inner speech developmentally. We address this question using behavioral (Experiment 1) and functional MRI (Experiment 2) approaches.

Experiment 1

Participants were given prompts and asked to silently generate a sentence or to create a visual image. Subsequently, they were asked to judge the vividness/clarity of the resulting sentence or image, or of the potentially accompanying representation in the other mode of thought (e.g., to judge the vividness of the visual image that may have accompanied the process of creating a sentence). If verbal and visual thinking are independent, then the clarity of the sentence should be high under the inner-speech instructions and low under the visual-imagery instructions, and the vividness of the image should be high under the visual-imagery instructions and low under the inner-speech instructions. However, if engaging in one mode of thought leads to the (involuntary) engagement in the other, then we would expect similar clarity/vividness ratings for that mode of thought regardless of the instructions.

We first conducted this experiment in a lab environment, and then replicated it online using Amazon.com's Mechanical Turk (AMT) marketplace.[1]

Lab version

Participants

Forty-two adults (twenty-two females), mean age=29.6 (standard deviation=10.2), from Harvard University and the surrounding community participated for pay or course credit. All participants gave informed consent in accordance with the Internal Review Board at Harvard.

Design and procedure

On each trial, participants were presented with two words on the screen: one denoted an occupation (e.g., *ballerina*), and the other denoted a location (e.g., *church*) or an inanimate object (e.g., *laptop*). The words for each category (64 occupations, 32 locations, and 32 objects) were selected randomly for each participant from a larger set of words, which included 159 occupations, 78 locations, and 79 objects (the full list of materials is available at https://evlab.mit.edu/papers/Amit_NI; the original intent was to have 160 occupations and 80 of each of locations and objects, but a few items were repeated accidentally). The 64 combinations of occupation-location and occupation-object words were random for each participant. Along with the two words, participants received an instruction to either engage in inner speech ("Create a sentence") or in visual imagery ("Imagine"; Fig. 1). In the inner-speech condition, participants were asked to silently generate an 8–10 word-long sentence that contains the two target words. In the visual-imagery condition, participants were asked to form an image of the person denoted by the occupation noun performing some action in the target location or with the target object. Before the experiment, participants were told that sometimes creating a sentence may be accompanied by some visual image(s), and forming a visual image may be accompanied by a "voice over" in one's head. They were told that during the experiment they would be asked to either judge the vividness of the representation they generated following the instructions on that trial, or of the potentially accompanying representation in the other mode of thought.

The prompt words and the instruction were presented for 900 ms, followed by a blank screen presented for 4,000 ms. Participants were asked to report the level of perceived vividness of the sentence or the image on a scale from 1 ("not at all vivid") to 7 ("very vivid"). Once a response was made, a fixation cross was presented for 800 ms, and then the next trial began.

The sixty-four trials were blocked by condition (i.e., create sentence/report vividness of sentence, create sentence/report vividness of image, imagine/report vividness of sentence, imagine/report vividness of image), with 16 trials per condition (8 occupation-location and 8 occupation-object trials, interleaved). Condition and trial order were random across participants, and there was a 3 s fixation period between each pair of blocks. The experiment took approximately 10 min.

Mechanical Turk version

Participants

381 adults (122 females, 114 males, 145 did not report gender), mean age=34.8 (standard deviation=12.6), recruited from the Amazon.com's Mechanical Turk marketplace, participated for payment. All participants gave informed consent in accordance with the Internal Review Board at Harvard.

Design and procedure

The design was similar to that of the lab version, with the following changes. First, in the lab experiment participants were asked to report the perceived vividness of both the sentences and the images. These instructions might be confusing when evaluating sentences because the notion of vividness is typically associated with visual images. Therefore, we adjusted the instructions asking the participants to report the perceived *clarity* of the sentence, on a scale from 1 ("not at all clear") to 7 ("very clear"). Second, unlike in the lab experiment, which used a within-subject design, a between-subjects design was used here. Each participant was randomly assigned to one of the four conditions, which led to 90–100 participants per condition (N=94 for the create sentence/report clarity of sentence condition, N=90 for the create sentence/report vividness of image condition, N=97 for the imagine/report clarity of sentence condition, and N=100 for the imagine/report vividness of image condition). Third, we increased the power by having each participant perform 160 trials. Finally, unlike in the lab experiment, where the occupation-location and occupation-object pairs were randomly generated for each participant, here we pre-generated two lists of 160 pairs (80 occupation-location and 80 occupation-object pairs; created from the set of 159 occupations, 78 locations, and 79 objects used in the lab experiment). Any given participant randomly received one of these lists. Trial order was the same within each list across participants. The experiment took approximately one hour.

Results

Lab version

Participants took 1,695 ms to respond on average. A 2 (instruction: create a sentence, imagine) x 2 (rating: vividness of the sentence, vividness of the image) repeated measures analysis of variance revealed a main effect of rating, with sentences being rated as more vivid than images (Ms=5 and 4.9, respectively; $F(1, 41)=4.19$, $p=.047$, eta squared=.09), and an interaction – which appears to be driving the main effect – between instruction and rating ($F(1, 41)=7.3$, $p=.01$). In particular, the vividness ratings of the sentences were higher in the "create a sentence" condition than in the "imagine" condition (Ms=5.2 and 4.8, respectively; $F(1, 41)=9.7$, $p=.003$, eta squared=.19). However, there was no difference in the vividness ratings of the images as a function of instruction (create a sentence, imagine; Ms=4.8 and 4.9, respectively; $F < 1$; Fig. 2).

[1] AMT is an online labor market where researchers recruit workers to complete tasks for payment. AMT has an advantage over lab experiments due to the recruitment of a more diverse population allowing for greater generalization (Horton et al., 2011). Recent evidence shows that experiments conducted originally in the lab replicate robustly on AMT (e.g., Buhrmester et al., 2011; Horton et al., 2011; Oppenheimer et al., 2009; Paolacci et al., 2010; Rand, 2012).

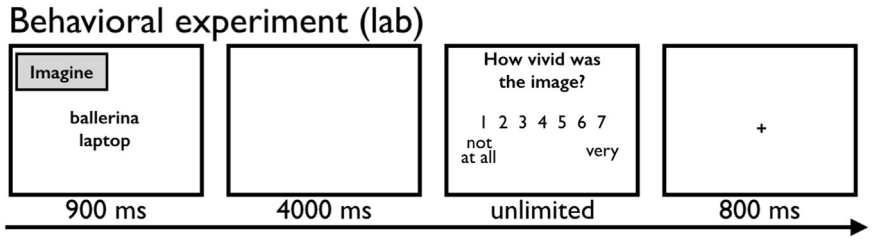

Fig. 1. A sample trial (from the imagine/report vividness of image condition) in the lab experiment.

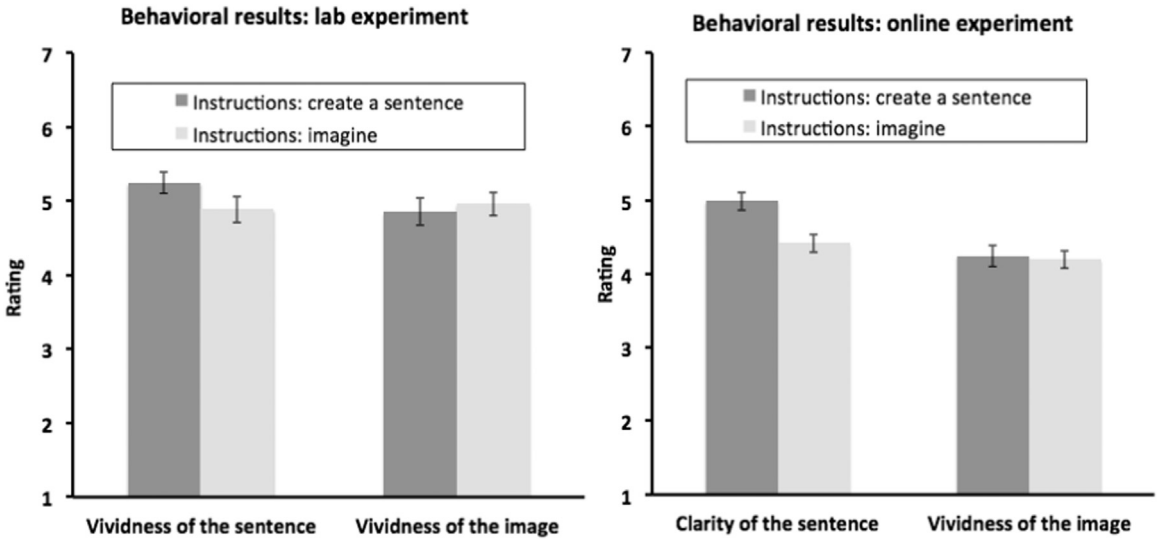

Fig. 2. Vividness ratings as a function of instruction (create a sentence, imagine) and the rated representation (sentence, image).

Mechanical Turk version

A 2 (instruction: create a sentence, imagine)×2 (rating: clarity of the sentence, vividness of the image) univariate analysis of variance revealed a main effect of rating, with sentences being rated as more clear than images (*Ms*=4.6 and 4.2, respectively; $F(1, 377)=13.9$, $p < .001$, eta squared=.03), and a main effect of instruction, with higher ratings in the "create a sentence" conditions than in the "imagine" conditions (*Ms*=4.6 and 4.3, respectively; $F(1, 377)=5.8$, $p=0.16$, eta squared=.01). However, as in the lab version, the main effects appear to be driven by the interaction between rating and instruction ($F(1, 377)=4.1$, $p=.04$). In particular, the clarity ratings of the sentences were higher in the "create a sentence" condition than in the "imagine" condition (*Ms*=4.9 and 4.4, respectively; $F(1, 377)=9.9$, $p=.002$, eta squared=.02). However, there was no difference in the vividness ratings of the images as a function of instruction (create a sentence, imagine; *Ms*=4.2 and 4.1, respectively; $F < 1$; Fig. 2).

Discussion

The results of the two versions of the behavioral experiment are similar and suggest that an asymmetry exists between inner speech and visual imagery: whereas the clarity of a verbal representation is higher when participants are generating a sentence compared to when they are attempting to form a visual image, the vividness of a visual image does not appear to be affected by whether participants are attempting to form an image versus generate a sentence. In other words, during verbal thought, a "by-product" visual image appears to be formed that is as vivid as the one generated when explicitly attempting to form an image. However, during visual thought, any verbal "voice-over" that is formed is significantly less robust than the one generated when explicitly attempting to think verbally. It is worth noting that the reported vividness of the images under the "imagine" instructions is lower than the reported clarity of the sentences under the "create-a-sentence" instructions in both experiments (lab: *Ms*=4.9 and 5.2, respectively, $t(41)=2.6$, $p=.01$; online: *Ms*=4.1 and 4.9, respectively, $t(192)=4.5$, $p < .001$), suggesting that visual representations may be generally less robust/vivid than the verbal ones. However, it is worth noting that visual imagery ratings fell reliably above the mid-point of the scale in both studies, ($t(41)=5.8$, $p < .0001$ in the lab version, and $t(189)=2.36$, $p < .019$ in the online version), suggesting that people do engage in the task. (This interpretation is further strengthened by the fMRI results below.).

Experiment 2

A possible limitation of the behavioral experiments is that they rely on the participants' introspection about their mental representations. Introspection may be challenging for abstract notions like modes of thought, and thus participants may not be able to accurately perform this assessment. We therefore turned to a more direct and implicit method – functional brain imaging – to examine the extent to which verbal vs. visual representations are active when we engage in verbal vs. visual thought. To do so, we examined the responses in brain regions that have been implicated in linguistic processing and visual processing while participants engaged in different forms of thought.

The neural mechanisms that support visual imagery have been extensively investigated (e.g., Ganis et al., 2004). The key finding from this body of work is that the very same brain regions that support visual perception are active during visual imagery (Ganis et al., 2004; O'Craven and Kanwisher, 2000; cf. General Discussion), albeit often to a lesser extent (e.g., O'Craven and Kanwisher, 2000). Although not many studies have explicitly investigated the brain mechanisms that support inner speech (cf. Shergill et al., 2001), research on language production speaks to this question, especially given that most fMRI studies of language production rely on covert production paradigms.

Such studies have shown that the brain regions of the language network – a set of frontal and temporal brain regions predominantly in the left hemisphere that are engaged during language comprehension – are active during covert production, both at the lexical/sub-lexical level (e.g., Smith et al., 1998; Poldrack et al., 1999; Burton et al., 2000; Indefrey and Levelt, 2004; Geva et al., 2011) and at the level of phrase and sentence production (e.g., Brown et al., 2006; Golestani et al., 2006; Menenti et al., 2012; McGuire et al., 1996b; McGuire et al., 1996a; Shergill et al., 2001). In summary, visual imagery appears to engage visual cortical regions, and inner speech (covert production) appears to engage the brain regions that support language understanding.

Here we adopt a functional localization approach where brain regions of interest are defined functionally in each individual brain, and their responses are then examined to the critical conditions of interest. Functional localizers benefit from higher sensitivity and functional resolution because they circumvent the well-documented inter-individual anatomical and functional variability (e.g., Nieto-Castañon and Fedorenko, 2012) while also minimizing the problem of "reverse inference" (e.g., Poldrack, 2011).

We used a language localizer task (Fedorenko et al., 2010) to identify the regions of the core language network, and a visual localizer task to identify regions engaged in high-level visual processing (specifically, in the processing of people's faces and bodies, since our materials involved humans, as described below). These two sets of functional regions of interest (fROIs) were then probed for their responses to inner speech and visual imagery in the critical experiment.

Following O'Craven and Kanwisher (2000), we used memory recall to probe thought processes. Participants were familiarized with a set of sentences and visual images prior to the scanning session. Then, during the scan, they either viewed those sentences and images, or – in the critical conditions – were asked to recall them from memory. If inner speech and visual imagery can be invoked independently of each other, then the language regions should respond selectively or preferentially during sentence recall, and the visual regions should respond selectively or preferentially during visual image recall. However, if an asymmetry exists between the two modes of thought, as our behavioral results suggest, then we might expect the language regions to be more active for recalling sentences than images, but the visual regions to show a similar level of response for recalling both images and sentences.

Participants

Eleven native English-speaking adults (6 females, mean age=25.7, standard deviation=4; 10 right-handed, 1 left-handed but with typical left-hemisphere lateralization for language) from MIT and the surrounding community participated for payment. The participants had normal or corrected-to-normal vision. All participants gave informed consent in accordance with the requirements of MIT's Committee on the Use of Humans as Experimental Subjects (COUHES). Three additional participants were scanned but excluded from the analyses: two because they turned out to not be native English speakers, and one because of poor activations during the visual localizer task (and consequently, the inability to define visual fROIs, as needed for the critical analyses). We conducted a post-hoc power analysis for the repeated measures ANOVA. Given 0.05 alpha, our sample size, and the effect size that we observed for the critical interaction effect, we had 0.99 power in our study.

Design and procedure

Each participant completed a language localizer task, a visual localizer task, and the critical inner speech and visual imagery task. Some participants further performed one or two unrelated experiments for other studies. The session lasted approximately two hours.

Language localizer

Participants read sentences (e.g., *The speech that the politician prepared was too long for the meeting*) and lists of nonwords (e.g., *Las tuping cusarists fick prell pront cre pome villpa olp wornetist cho*) in a blocked design. The *sentences > nonwords* contrast targets brain regions sensitive to high-level linguistic processing (Fedorenko et al., 2010, 2011).

Each stimulus consisted of 12 words/nonwords. For details of how the language materials were constructed, see Fedorenko et al. (2010). The materials are available for download at http://web.mit.edu/evelina9/www/funcloc.html. Stimuli were presented in the center of the screen, one word/nonword at a time, at the rate of 450 ms per word/nonword. Participants were prompted to read the stimuli and press a button at the end of each sequence when a hand icon appeared on the screen (for 400 ms). Additionally, a 100 ms duration blank screen appeared at the beginning of each trial and after the hand image, for a total trial duration of 6 s. The button press task was included to help participants stay awake and alert.

Experimental and fixation blocks lasted 18 s (with 3 trials per block) and 14 s, respectively. Each run (consisting of 5 fixation blocks and 16 experimental blocks, 8 per condition) lasted 358 s. Each participant completed 2 runs. Condition order was counterbalanced across runs.

Visual localizer

Participants viewed short movie clips of faces, bodies, scenes, objects, and scrambled objects (Julian et al., 2012; Pitcher et al., 2011) in a blocked design. The *faces > objects* contrast and the *bodies > objects* contrast target brain regions selectively engaged in the visual processing of faces and bodies, respectively (e.g., Downing et al., 2001; Kanwisher et al., 1997). Face- and body-selective regions have been shown to be robustly identifiable at the individual-subject level with this particular version of the localizer (Pitcher et al., 2011).

Each stimulus lasted 3 s. For details of how the materials were constructed, see Pitcher et al. (2011). Participants were instructed to watch the videos attentively.

Experimental and fixation blocks lasted 18 s (with 6 trials per block). Each run (consisting of 3 fixation blocks and 10 experimental blocks, 2 per condition) lasted 234 s. Each participant completed 4 runs. Condition order was counterbalanced across runs.

Critical task

Participants read sentences, viewed images, recalled sentences from memory, and recalled images from memory in a blocked design. The materials consisted of 120 sentence-image pairs. The images were diverse memorable color photographs of a person interacting with an easily identifiable object. The sentences were sentence-level descriptions of these photographs (see Fig. 3 for sample stimuli; the complete set of materials is available from https://evlab.mit.edu/papers/Amit_NI). These materials were distributed across two experimental lists following a Latin Square design (List 1: images from the odd-numbered items, sentences from the even-numbered items; List 2: images from the even-numbered items, sentences from the odd-numbered items). Any given participant saw either List 1 or List 2 (i.e., 60 images and 60 sentences), and thus either an image or a sentence version of an item (but not both). This was done in order to minimize "cross-coding", i.e., participants retrieving the corresponding verbal representations when viewing the images, and the corresponding image when reading the sentences.

A day before the fMRI scanning session, participants received the set of images and sentences (from List 1 or List 2) and were asked to try to memorize them as best as they could. Two hours prior to the scanning session, participants performed a behavioral training session. In the first part of the training session, they were presented with the images and sentences that they had received the day before, one at a time in random order. Critically, each image/sentence was presented

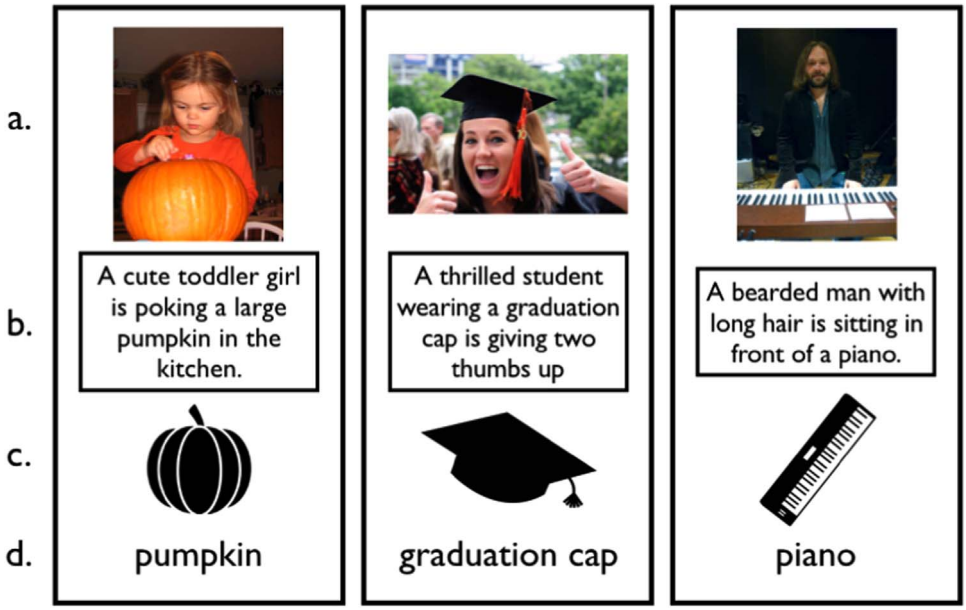

Fig. 3. Sample target images (a), target sentences (b), visual cues (c), and verbal cues (d). Target format was a within-subject variable; cue format was a between-subjects variable.

alongside a visual or verbal "cue" (the format of the cue varied across participants, such that any given participant saw only visual cues or only verbal cues). The cues were symbolic drawings (visual cues) or words (verbal cues) denoting the key object present/mentioned in the image/sentence (Fig. 3). The target images/sentences and the cues were presented side-by-side on the screen for 5 s. Participants were instructed to try to remember which cue was paired with which image/sentence because eventually they would be asked to recall the images/sentences when presented with the associated cues only. In the second part of the training, participants saw each cue for 5 s. They were asked to recall the corresponding image or sentence and to press the space bar when they were ready to continue. The task was self-paced. Once they pressed the space bar, they were shown the correct target for 5 s and asked to report whether the target matched the representation they recalled from memory by pressing "Y" for "yes", and "N" for "no". They were encouraged to answer truthfully. After each set of 20 trials, participants were offered a short break.

The goal of the training was to have participants reach 70% accuracy. Once this threshold was reached, the training ended. To ensure that participants were not answering "Y" without actually having recalled the target image/sentence, 12 foil trials were included. On these foil trials, participants were presented with cues (and targets) that they had not seen before and that therefore could not be associated with a target image/sentence. A "Y" answer on a foil trial would count as incorrect.

In the scanner, participants were presented with the images/sentences (in the perception conditions) or with the cues (in the recall conditions). When the targets were shown, participants were instructed to view the image or read the sentence; when the cue appeared, they were instructed to recall the corresponding image/sentence from memory.

Each trial lasted 3 s. The image/sentence or the cue appeared for 2.5 s, followed by a blank screen presented for 500 ms. Experimental and fixation blocks lasted 18 s (with 6 trials per block) and 14 s, respectively. Each run (consisting of 3 fixation blocks and 8 experimental blocks, 2 per condition) lasted 186 s. Each participant completed 5 runs. Condition order was counterbalanced across runs and participants.

fMRI data acquisition

Structural and functional data were collected on the whole-body 3 T Siemens Trio scanner at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. T1- weighted structural images were collected in 128 axial slices with 1.33mm isotropic voxels (TR=2,530 ms, TE=3.39 ms). Functional, blood oxygenation level dependent (BOLD), data were acquired using an EPI sequence (with a 90° flip angle and using GRAPPA with an acceleration factor of 2), with the following acquisition parameters: thirty-one 4 mm thick near-axial slices acquired in the interleaved order (with 10% distance factor), 2.1 mm×2.1 mm in-plane resolution, FoV in the phase encoding (A > > P) direction 200 mm and matrix size 96×96, TR=2000 ms and TE=30 ms. The first 10 s of each run were excluded to allow for steady state magnetization.

fMRI data preprocessing and analysis

MRI data were analyzed using SPM5 and custom Matlab scripts (available – in the form of an SPM toolbox – at http://web.mit.edu/evelina9/www/funcloc.html). Each participant's data were motion corrected and then normalized into a common brain space (the Montreal Neurological Institute, MNI template) and resampled into 2mm isotropic voxels. The data were then smoothed with a 4 mm Gaussian filter and high-pass filtered (at 200 s).

For all the key analyses, regions of interest were defined functionally in each participant. To define the individual fROIs, we used the Group-constrained Subject-Specific (GSS) approach developed in Fedorenko et al. (2010) and Julian et al. (2012). In particular, we intersected a set of functional "parcels" generated from group-level representations of the activations for each relevant localizer contrast (*sentences > nonwords*, *faces > objects*, and *bodies > objects*) in an independent group of participants with each individual participant's activation map for the same contrast. Voxels within each parcel were sorted based on their *t*-values, and the top 10% of voxels were chosen as that subject's functional region of interest.

Six language fROIs were defined in each participant: three on the lateral surface of the left frontal cortex (LIFGorb, LIFG, and LMFG fROIs) and three on the lateral surface of the temporal and parietal cortex (LAntTemp, LPostTemp and LAngG fROIs). These parcels, generated from a group-level representation of language localizer data

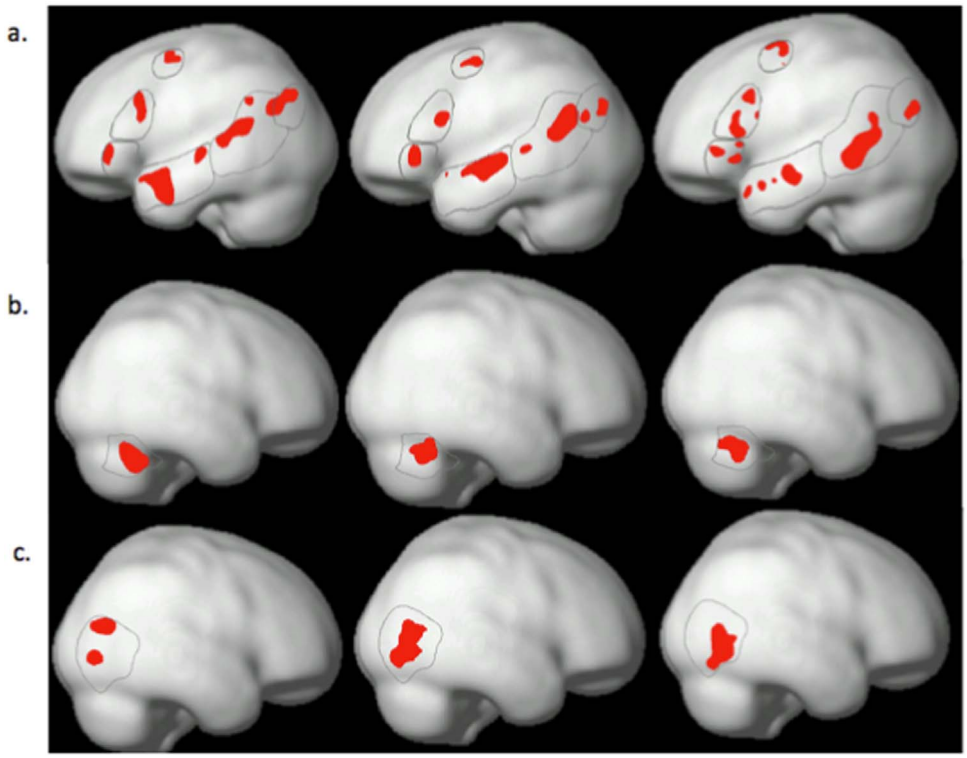

Fig. 4. a. Language fROIs in the left hemisphere of three typical subjects; b. Face fROIs (FFA) in the right hemisphere of three typical subjects; c. Body fROIs (EBA) in the right hemisphere of three typical subjects.

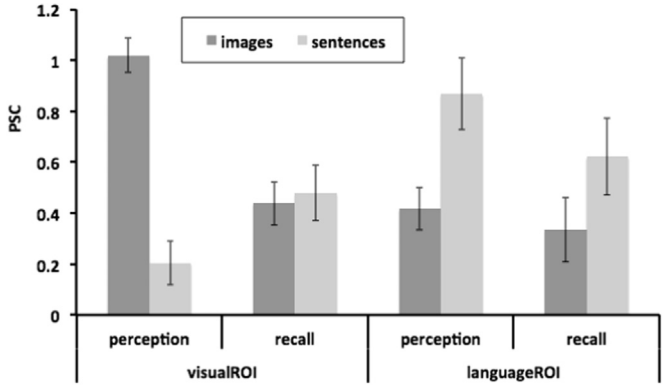

Fig. 5. Percent BOLD signal change (PSC) in the visual and language fROIs as a function of task (perception, recall) and target's format (image, sentence). Error bars represent standard error of the mean by participants.

from 220 participants, are similar to the parcels reported originally in Fedorenko et al. (2010) based on a set of 25 participants, except that the two anterior temporal parcels (LAntTemp, and LMidAntTemp) ended up being grouped into one, and so did the two posterior temporal parcels (LMidPostTemp and LPostTemp). The language parcels are available for download from http://web.mit.edu/evelina9/www/funcloc.html.

Four visual fROIs were defined in each participant: bilateral fusiform face area (FFA; Julian et al., 2012; Kanwisher et al., 1997) and bilateral extrastriate body area (EBA; Downing et al., 2001). The visual parcels are available for download from http://web.mit.edu/bcs/nklab/GSS.shtml. Sample language and visual fROIs are shown in Fig. 4.

To estimate the responses of the fROIs to the conditions of the critical experiment, all the data from the localizer experiments were used for defining the fROIs. To estimate the responses to the conditions of the localizer tasks (to ensure that fROIs behave as expected showing robust localizer responses), we used an across-runs cross-validation procedure, as described in Nieto-Castañon and Fedorenko (2012), so that the data used for fROI definition were independent from the data used for response estimation (Kriegeskorte et al., 2009). Statistical tests across participants were performed on the percent signal change (PSC) values extracted from the fROIs.

Results

Localizer experiments

As expected, all the fROIs showed the predicted behavior with respect to the responses to the localizer contrasts (estimated using data not used for fROI definition, as discussed above): all the language fROIs showed a robust *sentences* > *nonwords* effect ($ts > 9.3$, $ps < 0.0001$), the face-selective fROIs showed a robust *faces* > *objects* effect ($ts > 7.6$, $ps < 0.0001$), and the body-selective fROIs showed a robust *bodies* > *objects* effect ($ts > 12.5$, $ps < 0.0001$).

Critical experiment

The key results are shown in Fig. 5 and Table 1. For the main analysis, we pooled data from across all the language regions and all the visual regions because we had no reason to expect differences among regions (but see the results for each fROI separately in Table 1, which confirm qualitatively similar patterns across the language regions, and across the visual regions).

We first conducted a repeated measures analysis of variance, with cue format (visual, verbal) as a between-subjects variable, and task (perception, recall), target's format (image, sentence), and region of interest (visual fROIs, language fROIs) as within-subject factors. This analysis revealed no significant main effect of cue format ($F(1, 9) < 1$, n.s., eta squared=0.03), and no interactions between cue format and the target's format ($F(1, 9) < 1$, n.s., eta squared=0.01) or cue format and ROI ($F(1, 9)=1.2$, n.s., eta squared=0.11). The interaction between

Table 1
Effect of stimulus type in the visual and language fROIs for the perception (left column) and recall (right column) conditions. We report two-tailed t-tests. DF=10.

Region	**Perception conditions:**	**Recall conditions:**
	Images > sentences (for visual fROIs) / Sentences > images (for language fROIs)	Images > sentences (for visual fROIs) / Sentences > images (for language fROIs)
Averaging across the visual fROIs	t=9.8, $p < 0.0001$	$t < \|1\|$, *n.s.*
Averaging across the language fROIs	t=2.9, p=0.014	t=2.9, p=0.015
Visual-lFFA	t=1.8, p=0.09	$t < \|1\|$, *n.s.*
Visual-rFFA	t=9.7, $p < 0.0001$	$t < \|1\|$, *n.s.*
Visual - lEBA	t=6.23, $p < 0.0001$	$t < \|1\|$, *n.s.*
Visual-rEBA	t=11.76, $p < 0.0001$	t =−1.48, *n.s.*
Language - LIFGorb	t=2.01, p=0.07	t=2.6, p=0.02
Language - LIFG	t =3.7, p =.004	t=2.99, p=0.01
Language - LMFG	t =3.64, p=.004	t =2.78, p=0.02
Language - LAntTemp	t =1.9, p =.08	$t < \|1\|$, *n.s.*
Language - LPostTemp	t=3.62, p =.005	t =2.7, p=0.02
Language - LAngG	$t < \|1\|$, *n.s.*	$t < \|1\|$, *n.s.*

cue format and task (perception, recall) was significant (F(1, 9)=7.8, p=0.02, eta squared=0.46), such that the difference in activation between perception and recall was smaller when the cue was visual (Ms=0.54 and 0.48, respectively) than when the cue was verbal (Ms=0.73 and 0.44, respectively). However, the main effect of task, or its interaction with cue format, is not of central interest to the hypotheses evaluated here. As a result, we excluded cue format from further analyses.

A repeated measures analysis of variance with task (perception, recall), target's format (image, sentence), and region of interest (visual fROIs, language fROIs) as factors reveled several reliable effects and interactions. First, we observed a main effect of task, with stronger responses during the perception than the recall conditions (*Ms*=0.62 and 0.46, respectively; F(1, 10)=8.9, p < 0.02, eta squared=0.47). Second, we observed an interaction between the format of the target (image, sentence) and ROI (visual fROIs, language fROIs), such that the visual fROIs responded more strongly to images than sentences (*Ms*=0.73 and 0.34, respectively), whereas the language fROIs responded more strongly to sentences than images (*Ms*=0.74 and 0.37, respectively; F (1, 10)=40.1, p < .001, eta squared=0.78). Third, we observed an interaction between the format of the target (image, sentence) and task (perception, recall), such that in the perception conditions the response was stronger for images than sentences (*Ms*=0.71 and 0.53, respectively), but in the recall conditions the response was stronger for sentences than images (*Ms*=0.55 and 0.38, respectively; F (1, 10)=7.8, p < .02, eta squared=0.43). Critically, we observed a three-way interaction: F (1, 10)=26.8, p < .001, eta squared=0.72. The pattern of the results suggests the following interpretation: in the *visual fROIs*, the response was stronger for images than sentences in the perception conditions (*Ms*=1.02 and 0.2, respectively), but not in the recall conditions (*Ms*=0.43 and 0.47, respectively). In contrast, in the *language ROIs*, the response was stronger for sentences than images in the perception conditions (*Ms*=0.87 and 0.41, respectively), but *also* in the recall conditions (*Ms*=0.62 and 0.33, respectively). Thus, in line with what we had observed in our behavioral experiments, it appears that visual imagery is engaged to a similar extent regardless of whether individuals engage in visualizing or verbal thought, but verbal representations are invoked more strongly during verbal thought than during visual imagery.

Discussion

The results of the fMRI experiment converge with those of the behavioral experiment and suggest that an asymmetry exists between inner speech and visual imagery: people tend to generate visual images of what they think about verbally. However, the “voice-over” that people may generate when thinking visually is not nearly as strong / frequent as that generated during inner speech.

General discussion

We investigated the relationship between two modes of thought: inner speech and visual imagery, asking whether the two are independent. Converging evidence from two behavioral and one fMRI experiment suggests that an asymmetry exists between inner speech and visual imagery. In particular, individuals appear to have better control over inner speech: strong verbal representations are only invoked when individuals deliberately attempt to think verbally, but not when they engage in visual imagery. However, they generate similarly robust visual representations regardless of whether they attempt to engage in visual imagery or inner speech. In other words, visual representations appear to get invoked to a similar degree during both modes of thought.

Before discussing the implications of these results, it is worth noting that in addition to the perceptual visual brain regions mediating visual imagery (Ganis et al., 2004; O'Craven and Kanwisher, 2000) and high-level language processing brain regions mediating inner speech (e.g., Brown et al., 2006), there may be other brain regions that play a role in visual and/or verbal thought. The evidence comes primarily from patient studies. For example, some individuals with severe visual perception deficits can apparently nevertheless engage in visual imagery (e.g., Chatterjee and Southwood, 1995; Bartolomeo, 2002, 2008), and some individuals with intact perceptual abilities appear unable to engage in volitional imagery (e.g., Charcot and Bernard, 1883; Nielsen, 1946; Zeman et al., 2015), although most still experience involuntary imagery (Zeman et al., 2015). Similarly, some individuals with language perception / production deficits appear to be able to engage in inner speech (e.g., Hayward et al., 2014).

What are the brain regions – in addition to the visual perception regions and the language regions – that can support thought processes* Some have suggested that domain-general cognitive control regions of the fronto-parietal network may be important for visual imagery (e.g., Bartolomeo, 2008) and perhaps other kinds of imagery (McNorgan, 2012) and thought (e.g., Gerlach et al., 2011). Others have linked the default network (Buckner et al., 2008) to thought processes, especially those linked to our past experiences and future simulations (e.g., Buckner et al., 2008; Spreng and Grady, 2010; Andrews-Hanna et al., 2014). We did not investigate these brain networks in the current study, so our conclusions are limited to the role of visual perception regions and language regions in visual and verbal thought.

Our results have implications for any study that requires participants to think verbally or visually. The asymmetry we observed

suggests that people have varying degrees of volitional control over verbal vs. visual thinking. Thus, asking a participant to consider a verbal stimulus is likely to be accompanied by visual imagery. However, if the participant is asked to visually imagine something, this process is less likely to be accompanied by a "voice over."

Furthermore, given that the way we think affects downstream beliefs, emotions, and decisions (e.g., Amit and Greene, 2012), our findings may be relevant to the growing body of research on *prospection*, i.e., the way people think about the future (e.g., Seligman et al., 2013, 2016). For example, Amit et al. (2009) have shown that people use visual representations to prospect about proximal events, and verbal representations to prospect about distal events (see also Trope and Liberman, 2010). However, the asymmetrical relationship between inner speech and visual imagery observed here might suggest that visual representations are actually invoked when thinking about both proximal and distal events. This ubiquitous engagement of the visual system during thought processes a) may make distal events seem psychologically closer, and b) suggests that people are constrained, to some extent, to the "here and now". According to Amit et al. (2009), an efficient way to prospect about distal future is to think about it abstractly, using verbal means, given that verbal thought enables one to focus on the invariant gist and omit incidental details. However, our findings suggest that people may be unable to engage in verbal thought without invoking visual imagery. This property of our cognitive system may place a fundamental limitation on our prospection abilities, and on abstract thought in general.

Our findings might also relate to work on "cognitive styles", i.e., the preferred modes of thought that allegedly differ across individuals (e.g., Kirby et al., 1988; Kozhevnikov et al., 2005) and affect both the processing of external stimuli (e.g., Kraemer et al., 2014) and thinking during unconstrained cognition (Delamillieure et al., 2010). An interesting future direction would be to investigate how these stable individual preferences for verbal vs. visual thinking may interact with the asymmetry between the two modes of thought observed in the current study. In particular, do individuals with a preference for the verbal style perhaps have better control over the engagement of visual imagery*

As outlined in the Introduction, visual capacities emerge earlier than language both developmentally and evolutionarily, and may therefore be somehow more basic as a medium of thought, so that visual imagery is always present, to some degree, whereas inner speech is more under our volitional control. Interestingly, however, visual imagery appears to be less robust than verbal thought. For example, in our fMRI study, activation in the visual regions during visual imagery is only 35% of the response during actual visual perception, but activation in the language regions during inner speech is 60% of the response during linguistic processing, suggesting that the evoked verbal representations are more similar to actually experiencing the same stimuli externally compared to the evoked visual representations. Our results are further somewhat surprising in light of claims that visual imagery apparently requires substantial effort (e.g., Marschark and Cornoldi, 1991; Denis and Cocude, 1992) and may, in fact, be entirely absent in some individuals, as noted above. It is possible that the latter cases concern highly detailed visual imagery, whereas constructing coarse-level visual representations is fairly easy and, as our results suggest, present during both visual and verbal thought.

Author contributions

EA and EF designed the study and constructed the materials for the behavioral experiments. CH created the materials and script for the fMRI experiment. EA, CH, and NH collected the data. EA and EF analyzed and interpreted the data, and wrote the manuscript.

Acknowledgment

We thank Joshua D. Greene for helpful discussions, and Zach Mineroff for creating the website for this manuscript. This publication was made possible through the support of a grant from the John Templeton Foundation (37495), "Prospective Psychology Stage 2: A Research Competition" to Martin Seligman. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation. E.F. was supported by Eunice Kennedy Shriver NICHD Award R00 HD-057522.

References

Amit, E., Algom, D., Trope, Y., 2009. Distance-dependent processing of pictures and words. J. Exp. Psychol.: General. 138 (3), 400–415.
Amit, E., Greene, J.D., 2012. You see, the ends don't justify the means visual imagery and moral judgment. Psychol. Sci. 23 (8), 861–868.
Amit, E., Wakslak, C., Trope, Y., 2013. The use of visual and verbal means of communication across psychological distance. Pers. Social. Psychol. Bull. 39 (1), 43–56.
Andrews-Hanna, J.R., Smallwood, J., Spreng, R.N., 2014. The default network and self-generated thought: component processes, dynamic control, and clinical relevance. Ann. New Y. Acad. Sci. 1316 (1), 29–52.
Baddeley, A.D., Hitch, G., 1974. Working memory. Psychol. Learn. Motiv. 8, 47–89.
Barsalou, L.W., 2010. Grounded cognition: past, present, and future. Top. Cogn. Sci. 2 (4), 716–724.
Bartolomeo, P., 2002. The relationship between visual perception and visual mental imagery: a reappraisal of the neuropsychological evidence. Cortex 38 (3), *357–378*.
Bartolomeo, P., 2008. The neural correlates of visual mental imagery: an ongoing debate. Cortex 44 (2), 107–108.
Boroditsky, L., Prinz, J., 2008. What thoughts are made of. In: Semin, G., Smith, E. (Eds.), Embodied grounding: Social, cognitive, affective, and neuroscientific approaches.. Cambridge University Press, New York.
Brown, S., Martinez, M.J., Parsons, L.M., 2006. Music and language side by side in the brain: a PET study of the generation of melodies and sentences. Eur. J. Neurosci. 23, 2791–2803.
Buckner, R.L., Carroll, D.C., 2007. Self-projection and the brain. Trends Cogn. Sci. 11 (2), 49–57.
Buhrmester, M., Kwang, T., Gosling, S.D., 2011. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? Perspect. Psychol. Sci. 6 (1), 3–5.
Burton, M.W., Small, S.L., Blumstein, S.E., 2000. The role of segmentation in phonological processing: an fMRI investigation. J. Cogn. Neurosci. 12, 679–690.
Charcot, J.M., Bernard, D., 1883. Un cas de suppression brusque et isolée de la vision mentale des signes et des objets (formes et couleurs). Le. Prog. Méd. 11, 568–5571.
Carruthers, P., 2002. The cognitive functions of language. Behav. Brain Sci. 25 (6), 657–674.
Chatterjee, A., Southwood, M.H., 1995. Cortical blindness and visual imagery. Neurology 45 (12), 2189–2195.
Delamillieure, P., Doucet, G., Mazoyer, B., Turbelin, M.R., Delcroix, N., Mellet, E., Zago, L., Crivello, F., Petit, L., Tzourio-Mazoyer, N., Joliot, M., 2010. The resting state questionnaire: an introspective questionnaire for evaluation of inner experience during the conscious resting state. Brain Res. Bull. 81, 565–573.
Denis, M., Cocude, M., 1992. Structural properties of visual images constructed from poorly or well-structured verbal descriptions. Mem. Cogn. 20 (5), 497–506.
Dennett, D.C., 1991. Consciousness explained. Little, Brown and Co, Boston.
Downing, P.E., Jiang, Y., Shuman, M., Kanwisher, N., 2001. A cortical area selective for visual processing of the human body. Science 293 (5539), 2470–2473.
Fedorenko, E., Hsieh, P.J., Nieto-Castañón, A., Whitfield-Gabrieli, S., Kanwisher, N., 2010. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J. Neurophysiol. 104 (2), 1177–1194.
Frank, M., Fedorenko, E., Lai, P., Gibson, E., Saxe, R., 2012. Verbal interference suppresses exact numerical representation. Cogn. Psychol. 64, 74–92.
Ganis, G., Thompson, W.L., Kosslyn, S.M., 2004. Brain areas underlying visual mental imagery and visual perception: an fMRI study. Cogn. Brain Res. 20 (2), 226–241.
Gerlach, K.D., Spreng, R.N., Gilmore, A.W., Schacter, D.L., 2011. Solving future problems: default network and executive activity associated with goal-directed mental simulations. Neuroimage 55 (4), 1816–1824.
Geva, S., Jones, P.S., Crinion, J.T., Price, C.J., Baron, J.C., Warburton, E.A., 2011. The neural correlates of inner speech defined by voxel-based lesion–symptom mapping. Brain 134 (10), 3071–3082.
Golestani, N., Alario, F., Meriaux, S., Le Bihan, D., Dehaene, S., Pallier, C., 2006. Syntax production in bilinguals. Neuropsychologia 44 (7), 1029–1040.
Hayward, W., Fama, M.E., Sullivan, K.L., Snider, S.F., Lacey, E.H., Friedman, R.B., Turkeltaub, P.E., 2014. Inner speech in people with aphasia. In Frontiers in Psychology Conference Abstract: Academy of Aphasia–In: Proceedings of the 52nd Annual Meeting.
Horton, J.J., Rand, D.G., Zeckhauser, R.J., 2011. The online laboratory: conducting experiments in a real labor market. Exp. Econ. 14 (3), 399–425.
Hume, D., 1739/1951. In: Selby-Bigge, L.A. (Ed.), Enquiry: A treatise of human nature. Clarendon Press, Oxford, England.
Indefrey, P., Levelt, W.J., 2004. The spatial and temporal signatures of word production components. Cognition 92 (1), 101–144.

Julian, J., B., Fedorenko, E., Webster, J., Kanwisher, N., 2012. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. Neuroimage 60 (4), 2357–2364.
Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17 (11), 4302–4311.
Kirby, J.R., Moore, P.J., Schofield, N.J., 1988. Verbal and visual learning styles. Contemp. Educ. Psychol. 13, 169–184.
Kozhevnikov, M., Kosslyn, S., Shephard, J., 2005. Spatial versus object visualizers: a new characterization of visual cognitive style. Mem. Cogn. 33 (4), 710–726.
Kraemer, D.J., Hamilton, R.H., Messing, S.B., DeSantis, J.H., Thompson-Schill, S.L., 2014. Cognitive style, cortical stimulation, and the conversion hypothesis. Front. Human. Neurosci. *8*, 1–9, (Golestani).
Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12 (5), 535–540.
Marschark, M., Cornoldi, C., 1991. Imagery and verbal memoryImagery and cognition. Springer, US, 133–182.
McGuire, P.K., Silbersweig, D.A., Frith, C.D., 1996. Functional neuroanatomy of verbal self-monitoring. Brain 119 (3), 907–917.
McGuire, P.K., Silbersweig, D.A., Murray, R.M., David, A.S., Frackowiak, R.S.J., Frith, C.D., 1996. Functional anatomy of inner speech and auditory verbal imagery. Psychol. Med. 26 (1), 29–38.
McNorgan, C., 2012. A meta-analytic review of multisensory imagery identifies the neural correlates of modality-specific and modality-general imagery. Front. Human. Neurosci. 6, 285.
Menenti, L., Segaert, K., Hagoort, P., 2012. The neuronal infrastructure of speaking. Brain Lang. 122 (2), 71–80.
Morin, A., Michaud, J., 2007. Self-awareness and the left inferior frontal gyrus: inner speech use during self-related processing. Brain Res. Bull. 74 (6), 387–396.
Nielsen, J.M., 1946. Agnosia, apraxia, aphasia: their value in cerebral localization. PB Hoeber, Incorporated, New York, NY.
Nieto-Castañón, A., Fedorenko, E., 2012. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. Neuroimage 63 (3), 1646–1669.
O'Craven, K.M., Kanwisher, N., 2000. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. J. Cogn. Neurosci. 12 (6), 1013–1023.
Oppenheimer, D.M., Meyvis, T., Davidenko, N., 2009. Instructional manipulation checks: detecting satisficing to increase statistical power. J. Exp. Soc. Psychol. 45, 867–872.
Paolacci, G., Chandler, J., Ipeirotis, P.G., 2010. Running experiments on amazon mechanical turk. Judgm. Decis. Mak. 5 (5).
Paivio, A., 1986. Mental representations a dual coding approach. Oxford University Press, New York.
Pitcher, D., Dilks, D., Saxe, R., Triantafyllou, C., Kanwisher, N., 2011. Differential selectivity for dynamic versus static information in face-selective cortical regions. Neuroimage. 56 (15), 2356–2363.
Pléh, C., 2002. Speech as an opportunistic vehicle of thinking. Behav. Brain Sci. 25 (6), 695–696.
Poldrack, R.A., 2011. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron 72 (5), 692–697.
Poldrack, R.A., Wagner, A.D., Prull, M.W., Desmond, J.E., Glover, G.H., Gabrieli, J.D., 1999. Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. Neuroimage 10 (1), 15–35.
Price, C.J., 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage 62, 816–847.
Rand, D.G., Greene, J.D., Nowak, M.A., 2012. Spontaneous giving and calculated greed. Nature 489 (7416), 427–430.
Schrauf, R.W., 2002. Bilingual inner speech as the medium of cross-modular retrieval in autobiographical memory. Behav. Brain Sci. 25 (06), 698–699.
Seligman, M.E., Railton, P., Baumeister, R.F., Sripada, C., 2016. Homo Prospectus. Oxford University Press.
Seligman, M.E.P., Railton, P., Baumeister, R., Sripada, C., 2013. Navigating into the future or driven by the past. Perspect. Psychol. Sci. 8 (2), 119–141.
Shepard, R.N., Metzler, J., 1971. Mental rotation of three-dimensional objects. Science 171, 701–703.
Shergill, S.S., Bullmore, E.T., Brammer, M.J., Williams, S.C.R., Murray, R.M., McGuire, P.K., 2001. A functional study of auditory verbal imagery. Psychol. Med. 31 (2), 241–253.
Siegrist, M., 1995. Inner speech as a cognitive process mediating self-consciousness and inhibiting self-deception. Psychol. Rep. 76 (1), 259–265.
Smith, E.E., Jonides, J., Marshuetz, C., Koeppe, R.A., 1998. Components of verbal working memory: evidence from neuroimaging. Proc. Natl. Acad. Sci. USA 95, 876–882.
Sokolov, A.N., 1972. Inner Speech and Thought, translated by George T. Onischenko.
Spreng, R.N., Grady, C.L., 2010. Patterns of brain activity supporting autobiographical memory, prospection, and theory of mind, and their relationship to the default mode network. J. Cogn. Neurosci. 22 (6), 1112–1123.
Trope, Y., Liberman, N., 2010. Construal-level theory of psychological distance. Psychol. Rev. 117 (2), 440–463.
Vygotsky, L.S., 1962. Thought and Language. MIT Press, Cambridge, MA.
Vygotsky, L.S., 2012. Thought and Language. MIT press, Cambridge, MA.
Zeman, A., Dewar, M., Della Sala, S., 2015. Lives without imagery–Congenital aphantasia. Cortex, 378–380.
Zivin, G., 1979. The Development of self-regulation through private speech. Wiley, New York.