NeuroImage: Clinical 16 (2017) 399–408

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Validity and reliability of four language mapping paradigms

Stephen M. Wilson[a,b,c,*], Alexa Bautista[a], Melodie Yen[a,c], Stefanie Lauderdale[a], Dana K. Eriksson[a]

[a]Department of Speech, Language, and Hearing Sciences, University of Arizona, Tucson, AZ, USA
[b]Department of Neurology, University of Arizona, Tucson, AZ, USA
[c]Department of Linguistics, University of Arizona, Tucson, AZ, USA

ARTICLE INFO

Article history:
Received 9 February 2016
Received in revised form 1 March 2016
Accepted 20 March 2016
Available online 24 March 2016

Keywords:
Language mapping
fMRI
Validity
Reliability
Test-retest reproducibility

ABSTRACT

Language areas of the brain can be mapped in individual participants with functional MRI. We investigated the validity and reliability of four language mapping paradigms that may be appropriate for individuals with acquired aphasia: sentence completion, picture naming, naturalistic comprehension, and narrative comprehension. Five neurologically normal older adults were scanned on each of the four paradigms on four separate occasions. Validity was assessed in terms of whether activation patterns reflected the known typical organization of language regions, that is, lateralization to the left hemisphere, and involvement of the left inferior frontal gyrus and the left middle and/or superior temporal gyri. Reliability (test-retest reproducibility) was quantified in terms of the Dice coefficient of similarity, which measures overlap of activations across time points. We explored the impact of different absolute and relative voxelwise thresholds, a range of cluster size cutoffs, and limitation of analyses to a priori potential language regions. We found that the narrative comprehension and sentence completion paradigms offered the best balance of validity and reliability. However, even with optimal combinations of analysis parameters, there were many scans on which known features of typical language organization were not demonstrated, and test-retest reproducibility was only moderate for realistic parameter choices. These limitations in terms of validity and reliability may constitute significant limitations for many clinical or research applications that depend on identifying language regions in individual participants.

1. Introduction

Language areas of the brain can be mapped with functional magnetic resonance imaging (fMRI), typically by contrasting conditions that involve language processing with conditions that do not (Binder et al., 2008). Language localization varies in individuals, not only in terms of lateralization but also in terms of the specific regions that are involved in language processing within the dominant and non-dominant hemispheres. Variability is particularly pronounced in neurological patients in whom typical language regions may be damaged or dysfunctional (Berl et al., 2014), but language localization is variable in neurologically normal people as well (Knecht et al., 2003; Tzourio-Mazoyer et al., 2010; Seghier et al., 2011).

In several clinical and research contexts, such as pre-surgical functional mapping (Binder et al., 2008) and longitudinal studies of recovery from aphasia (Meltzer et al., 2009; Kiran et al., 2013), it is important to determine language localization in individual participants. In these contexts, the validity and reliability of language maps are critically important. Validity refers to the extent to which all and only the regions actually critical for language processing are identified as such. Reliability is the extent to which a language map is reproducible in the same participant on a different occasion. This can also be referred to as test-retest reproducibility.

The aim of this study was to investigate the validity and reliability of four language mapping paradigms, in neurologically normal older adults. Our specific goal was to investigate paradigms that might be effective for language mapping in longitudinal studies of individuals with aphasia. There are several important differences between language mapping in people with aphasia and language mapping in pre-surgical patients, who do not typically have significant language deficits (Binder et al., 2008). First, because language function is disrupted in aphasia, it is desirable to record behavioral responses in the scanner to ensure that the task(s) are being carried out (Thompson and den Ouden, 2008; Meinzer et al., 2013). In contrast, covert responses are typically used in pre-surgical contexts because they involve less head movement and speech-related artifacts, and have been empirically demonstrated to be psychometrically superior (Partovi et al., 2012). Second, since people with aphasia often make errors in performing language tasks, it is desirable to be able to separate correct and incorrect trials, since these differ in terms of their neural correlates (Fridriksson et al., 2009; Postman-Caucheteux et al., 2010). Separating trials generally requires event-related designs. In contrast, block designs are

* Corresponding author at: Department of Speech, Language, and Hearing Sciences, University of Arizona, P.O. Box 210071, Tucson, AZ 85721, USA.
E-mail address: smwilson@u.arizona.edu (S.M. Wilson).

http://dx.doi.org/10.1016/j.nicl.2016.03.015

typically used in pre-surgical contexts, since patients make few errors, and block designs offer greater power. Third, if language areas are to be identified in patients with moderate to severe aphasia, very simple paradigms may be required (Price et al., 2006). Conventional control conditions can be confusing to some neurological patients because of the task-switching demands they entail. In contrast, pre-surgical patients are generally capable of performing complex tasks which more thoroughly engage the language system (Binder et al., 2008, Binder et al., 2011). Of the four paradigms we investigated, two were expressive—sentence completion and picture naming—and involved overt responses and event-related designs in which correct and incorrect responses could be separated. Two were receptive—narrative comprehension and naturalistic comprehension—and involved comprehension of an audiobook and an edited television show, requiring no responses, thus making them suitable for more impaired patients. Across the four paradigms, scan time and analytical procedures were kept constant so that the paradigms could be compared.

The validity of language mapping paradigms has been investigated in several different ways. Concurrent validity has been examined by comparing fMRI to the Wada test (intracarotid amobarbital test) for language lateralization, and to electrocortical stimulation mapping (ESM) for language localization within a hemisphere. Concordance of lateralization between fMRI and Wada is generally high (Binder et al., 1996; Woermann et al., 2003; Benke et al., 2006; Janecek et al., 2013b; see Bauer et al., 2014 for review). While the Wada test is often considered a "gold standard" for lateralization of language function, it is not always reliable (Kho et al., 2005; Lanzenberger et al., 2005), and a recent investigation of a small sample of discordant cases showed that postoperative deficits were better predicted by fMRI than Wada in a majority of patients (Janecek et al., 2013a). Concordance of within-hemisphere localization between fMRI and ESM is moderate, and highly variable depending on methodological details (Yetkin et al., 1997; Rutten et al., 2002a; Pouratian et al., 2002; Bizzi et al., 2008; see Giussani et al., 2010 for review). Like the Wada test, ESM is not necessarily infallible as a "ground truth": stimulation is limited to the exposed surfaces of gyri, and language areas are not identified at all in a significant minority of patients (Sanai et al., 2008). Moreover, it is not known whether all cortical areas identified as critical for language by ESM are actually indispensable, because most language deficits subsequent to resective surgery resolve rapidly (Penfield and Roberts, 1959; Wilson et al., 2015). Ultimately the simple concept of "eloquent cortex" is limited: while there are undoubtedly specific brain regions that are important for language, their functions can often be compensated to varying degrees, and damage to numerous motor, perceptual, cognitive, attentional, and executive networks can impact language production or comprehension, even though these networks are not language-specific.

Another less direct approach to assessment of validity has been to determine whether language mapping paradigms activate left-lateralized frontal and temporal regions in neurologically normal individuals (Rutten et al., 2002b; Seghier et al., 2004; Harrington et al., 2006; Binder et al., 2008). Since the concept of left-lateralized frontal and temporal language regions is firmly established, this can be seen as a test of construct validity. In our study, we investigated healthy participants, so we had no external sources of information regarding language localization, therefore this was the approach we took. Specifically, we calculated lateralization indices to determine whether activations were more extensive in the left hemisphere, and we determined how frequently each paradigm activated left frontal and left temporal regions. While there are clear limitations to this approach, for a language mapping paradigm to be valid, it is necessary but not sufficient that it produces left-lateralized activation of inferior frontal and superior and middle temporal regions in most healthy participants, which is the predominant pattern in adults of all ages (Knecht et al., 2003; Tzourio-Mazoyer et al., 2010; Seghier et al., 2011).

Reliability, or test-retest reproducibility, is also important. Indeed, reliability places an upper limit on validity. In studies of recovery from aphasia, it is generally believed that recovery depends on neuroplasticity, that is, functional reorganization of language processing regions over time. Investigating neuroplasticity requires being able to distinguish genuine changes from scan-to-scan variability. Reliability of fMRI paradigms is generally assessed by having the same participants perform the task two or more times, and then calculating a similarity metric between the activations obtained each time (Bennett and Miller, 2010). A commonly used similarity metric is the intraclass correlation coefficient (ICC), which has been used in several reliability studies of language mapping paradigms (Fernández et al., 2003; Eaton et al., 2008; Meltzer et al., 2009). However the ICC does not provide a global measure of activation similarity; it must be performed on a specified region of interest, or voxel-by-voxel. Moreover, the ICC is calculated by dividing variance between subjects by total variance (between and within subjects), so it does not quantify test-reproducibility in any individual except with reference to a defined group.

In our view, a more useful similarity metric is the Dice coefficient of similarity, which was first used in neuroimaging by Rombouts et al. (1997). The Dice coefficient is a measure of the extent of overlap between thresholded activation maps obtained on two or more occasions, and is calculated as follows (for two sessions):

$$\text{Dice coefficient} = 2.V_{overlap}/(V_1 + V_2).$$

$V_{overlap}$ is the number of overlapping voxels, V_1 is the number of voxels activated at time 1, and V_2 is the number of voxels activated at time 2. If there are more than two sessions, the Dice coefficient can be averaged across all pairwise comparisons between sessions. A Dice coefficient of 0 implies no overlap at all between activations, whereas a Dice coefficient of 1 implies perfect overlap. A Dice coefficient between 0 and 1 can be interpreted intuitively as the probability that an activated voxel in one session will be activated in the other session. The Dice coefficient can be calculated based on activations over the whole brain, or can be restricted to activations in smaller regions of interest, such as the set of all potential language regions in both hemispheres, or even just a single region such as the inferior frontal gyrus. The advantages of the Dice coefficient are that it is widely used, it can be calculated in any individual without reference to a group, it yields a single metric of overall activation similarity encompassing all brain regions under consideration, and it is intuitive and easy to interpret (Bennett and Miller, 2010).

There is a substantial literature on test-retest reproducibility of language paradigms (Brannen et al., 2001; Maldjian et al., 2002; Rutten et al., 2002b; Fernández et al., 2003; Billingsley-Marshall et al., 2004; Harrington et al., 2006; Rau et al., 2007; Fesl et al., 2010; Gonzalez-Castillo and Talavage, 2011; Maïza et al., 2011; Gross and Binder, 2014). Many of these papers have reported Dice coefficients or related metrics, but to our knowledge, no previous studies have reported Dice coefficients except in specific regions of interest for any of the four paradigms we investigated in this study.

2. Methods

2.1. Participants

Five healthy older adults (aged 70–76 years; 3 females) took part in the study. The participants were recruited after attending a talk on language and the brain at a community center. They were all right-handed, native speakers of English, and neurologically normal. Their scores on the Mini Mental State Examination (Folstein et al., 1975) ranged from 29 to 30, and they were all at or near ceiling on an in-house aphasia battery. All participants gave written informed consent, and were modestly compensated for their time. The study was approved by the Institutional Review Board at the University of Arizona.

One additional participant consented to participate in the study, but was excluded due to hearing loss, which prevented her from hearing the stimuli over the scanner noise.

2.2. Neuroimaging protocol

Participants were scanned on four separate occasions. Sessions for each participant were separated by a mean of 22.5 days (range 12–42).

MRI data were acquired on a Siemens Skyra 3T scanner with a 20-channel head coil. Auditory stimuli were presented using insert earphones (S14, Sensimetrics, Malden, MA) padded with foam to attenuate scanner noise and reduce head movement. The presentation volume was adjusted to a comfortable level for each participant. Visual stimuli were presented on a 24″ MRI-compatible LCD monitor (BOLDscreen, Cambridge Research Systems, Rochester, UK) positioned at the end of the bore, which participants viewed through a mirror mounted to the head coil. Auditory and visual stimuli were controlled with the Psychophysics Toolbox version 3.0.10 (Brainard, 1997; Pelli, 1997) running under MATLAB R2012b (Mathworks, Natick, MA) on a Lenovo S30 workstation.

Each session included four language paradigms. For three of the four paradigms, $T2^*$-weighted BOLD echo planar images were collected with the following parameters: 210 volumes; 29 axial slices in ascending order; slice thickness = 3.5 mm with a 0.9 mm skip; field of view = 214 × 240 mm; matrix = 82 × 92 mm; repetition time (TR) = 2000 ms; echo time (TE) = 30 ms; flip angle = 90°; voxel size = 2.6 × 2.6 × 3.5 mm. For the fourth paradigm, the parameters were the same except that a sparse sampling sequence was used to acquire 45 volumes with TR = 9500 ms; acquisition time (TA) = 2000 ms. Initial dummy scans were also acquired (2 or 1 respectively) to allow magnetization to reach steady state. All functional series were exactly 7 min in duration, not including the dummy scans.

For anatomical reference, T1-weighted MPRAGE structural images were also acquired (voxel size = 0.9 × 0.9 × 0.9 mm).

2.3. Language mapping paradigms

The four language mapping paradigms were not identical in their structure, but were intended to make the best possible use of 7 min of scan time, depending on the particular goals of the paradigm. Each paradigm had four versions with different items, and the orders of the four paradigms within sessions, and the four versions across sessions, were counterbalanced across participants with one exception noted below.

In the *sentence completion* paradigm, participants heard or read a sentence that was missing its final word, and produced the final word out loud. This paradigm combines receptive language processing (understanding the sentence) and expressive language processing (producing the missing word). A rapid event-related design was used so that correct and incorrect trials could be separated. Twenty of the trials were presented auditorily and twenty were presented visually. To reduce task-switching demands for language-impaired patients, there was no control condition; instead a conjunction between the auditory and visual conditions was used to eliminate activations related to modality-specific sensory processing.

The stimuli were selected from those normed by Block and Baldwin (2010). Higher cloze sentences were selected, such that the mean cloze probability was 0.92 (range 0.87–0.99). The mean sentence length was 8.0 words (range 5–12 words). The cloze probabilities and sentence lengths were matched across the four versions of the paradigm. The sentences were recorded in a soundproof booth by a female, with rising intonation to cue each missing final word. The recorded sentences had a mean duration of 3559 ± 549 ms (SD) (range 2229–4795 ms). Written sentences were displayed for 4.5 s. The mean inter-trial interval (from onset to onset) was 10.2 s (range 6–20 s). This was longer than is typical in a rapid event-related design, in order to allow sufficient time to present auditory sentences, and to allow for relatively long response times, since the task was ultimately intended for individuals with aphasia. In between trials, and during auditory trials, participants fixated on a central crosshair.

The *picture naming* paradigm required participants to name pictures out loud. This paradigm focuses on expressive language processing, and has been widely used in studies of recovery from aphasia (e.g. Eaton et al., 2008; Fridriksson et al., 2009; Abel et al., 2015), in part because anomia is a ubiquitous feature of all aphasia subtypes. A rapid event-related design was again used so that correct and incorrect trials could be separated. There were 52 trials. To reduce task-switching demands for language-impaired patients, there was no control condition (as in Abel et al., 2015). This was considered feasible because the occipito-temporal regions activated by typical control conditions such as processing scrambled objects are well established (e.g. Wilson et al., 2009) and can readily be excluded from analysis.

The stimuli were colorized versions (Rossion and Pourtois, 2004) of the Snodgrass and Vanderwart (1980) pictures. Items with multimorphemic targets, or with name agreement <60% were not used. The mean length of the target names was 4.5 ± 1.5 (SD) phonemes (range 3–9), the mean log frequency of the targets based on the HAL corpus (Lund and Burgess, 1996) was 8.6 ± 1.5 (SD) (range 4.7–12.2), and the mean name agreement was 90.4 ± 9.7% (SD) (range 60–100%) The means and distributions of the numbers of phonemes and frequencies were matched across the four versions of the paradigm. Each picture was displayed for 3 s. The mean inter-trial interval (from the onset of one trial to the onset of the next trial) was 7.9 ± 2.7 s (SD) (range 5–16 s). In between trials, participants fixated on a central crosshair.

In the *naturalistic comprehension* paradigm, participants simply viewed a 7-min edited television program. This paradigm involves receptive language processing, and was considered to be a task that even the most impaired patients would likely be able to perform. Language regions were identified by comparing neural responses during periods when characters were speaking to periods when they were not.

Each of the four versions was derived from a different episode of the television series *Freaks and Geeks* (1999–2000). This series was chosen because it is engaging and well-acted, yet unfamiliar to most people, especially older people. One of the interweaved storylines from each episode was selected, then edited so as to create a coherent story that was 7 minute long or slightly longer (but MRI data acquisition ended at exactly 7 min), and such that speech took place approximately half of the time. Potentially offensive language, drug references, and sexual references were edited out. The mean proportion of language in the four videos was 51.1 ± 1.4% (SD) (range 49.8–52.2%).

In the *narrative comprehension* paradigm, participants listened to an audiobook as well as segments of reversed speech and silence. Like the naturalistic paradigm, this paradigm involves receptive language processing, and similar paradigms have been widely used in studies of people with aphasia (e.g. Crinion and Price, 2005). This was the only paradigm in which a sparse sampling sequence was used, because sparse sampling can be advantageous in studies of older adults, many of whom have some degree of hearing loss, and this was the only paradigm for which the temporal structure lent itself to a sparse sampling approach. There were 17 segments of narrative speech, 17 segments of reversed speech, and 10 silent intervals. One initial image was acquired, and then one image was acquired after each stimulus (or silent interval), for a total of 45 images. The narrative or backwards narrative segments were centered in these silent intervals, such that the peak of a typical hemodynamic response to each segment would coincide with acquisition of the subsequent image.

The narrative was the beginning of an audiobook recording of the novel *Hope Was Here* by Joan Bauer, read by Jenna Lamia (Bauer, 2004). The narrative was split into segments at pauses such that each segment was as long as possible up to 7 s (occasionally, slightly longer segments were extracted, then reduced to 7 s by shortening internal pauses). The segments had a mean duration of 5912 ms ± 978 (SD) ms (range 3123–7000 ms). The narrative, backwards, and silent segments were presented in pseudorandom order, but arranged in blocks as follows: The 17 narrative segments were presented in 5 blocks of 3,

plus one block of 2. Similarly, the 17 backwards segments were presented in 5 blocks of 3, plus one block of 2. The 10 silent intervals were presented in 5 blocks of 2. Unlike the other three paradigms, the four versions of the narrative comprehension paradigm were always presented in the same order across the four sessions, such that the narrative progressed from each session to the next.

2.4. Analysis of neuroimaging data

The functional data were first preprocessed with tools from AFNI version 2011_12_21_1014 (Cox, 1996). Head motion was corrected, with six translation and rotation parameters saved for use as covariates. There was modestly more head motion in the paradigms that involved overt speech (rotation: sentence completion: 0.072°/volume, picture naming: 0.075°/volume, naturalistic comprehension: 0.044°/volume, narrative comprehension: 0.058°/volume; translation: sentence completion: 0.13 mm/volume, picture naming: 0.13 mm/volume, naturalistic comprehension: 0.08 mm/volume; narrative comprehension 0.10 mm/volume; $p < 0.05$, Tukey's HSD). Next, the data were detrended with a Legendre polynomial of degree 2, and smoothed with a Gaussian kernel (FWHM = 6 mm). Then, independent component analysis (ICA) was performed using the fsl tool *melodic* version 3.14 (Beckmann & Smith, 2004). Noise components were manually identified with reference to the criteria of Kelly et al. (2010) and removed using *fsl_regfilt*. With one exception noted below, task models were convolved with a hemodynamic response function (HRF) based on a gamma density function (time to peak = 5.4 s, FWHM = 5.2 s), and fit to the data with the program *fmrilm* from the FMRISTAT package (Worsley et al., 2002). The six motion parameters were included as covariates, as were time-series from white matter and CSF regions (means of voxels segmented as white matter or CSF in the vicinity of the lateral ventricles) and three cubic spline temporal trends. The T1-weighted anatomical images were warped to MNI space using unified segmentation in SPM5 (Ashburner & Friston, 2005). Functional images were coregistered with structural images, and warped to MNI space.

In the *sentence completion* paradigm, trials were considered correct when participants made a single correct response of one or more words. Correct trials were modeled with two explanatory variables: one for auditory presentation and one for visual presentation. Each item was modeled as beginning at the onset of presentation, and having a duration of 4 s, and convolved with the HRF. All other trials, namely incorrect responses, multiple responses, or trials with no response, were modeled with two separate explanatory variables—one for auditory presentations and one for visual presentations—convolved with the HRF, and were not examined further. Finally, the voxelwise minimum was taken of the contrasts of correct responses to auditory items relative to rest, and correct responses to visual items relative to rest.

The *picture naming* paradigm was analyzed in a similar manner. Correct trials were modeled with a single explanatory variable, with onsets at the time of picture presentation, and event duration of 2 s. Incorrect responses, multiple responses, and trials without responses were modeled with a separate variable. The correct trials variable was contrasted with the resting baseline.

The *naturalistic comprehension* paradigm was analyzed using an explanatory variable that encoded the presence of language in the videos. The segments containing language were contrasted with segments not containing language (the latter comprising the baseline). As described above, characters were speaking during approximately half of each video. In order to account for acoustic differences between segments containing language and those not containing language, a second explanatory variable was used to account for auditory activity: root mean square acoustic energy was calculated in bins of 200 ms. Both variables were convolved with the HRF.

For the *narrative comprehension* paradigm, no HRF was modeled; instead, each volume was assumed to reflect the neural response to the immediately preceding segment. Segments of narrative were contrasted to segments of reversed narrative; silent segments were not used in the analysis.

2.5. Measures of validity and reliability

To assess validity, lateralization indices (LIs) were calculated according to the standard formula:

$$LI = (V_{Left} - V_{Right})/(V_{Left} + V_{Right}).$$

For each participant and paradigm, the LI was averaged across the four sessions. Note that if there are no activated voxels in either hemisphere, then LI is undefined. In those cases, LI was averaged across the other sessions, excluding the session(s) with no activated voxels, and this limitation was indicated.

The proportion of the 20 scans (4 scans for each of 5 participants) in which left frontal and left temporal language regions were activated were calculated. The left frontal region was defined as the pars opercularis, triangularis and orbitalis of the left IFG. The left temporal region was defined as the left MTG and the part of the left STG that was within 8 mm of the MTG. These regions were deemed activated when at least 50 voxels were activated.

For each participant and paradigm, a Dice coefficient of similarity was calculated as the mean of the six pairwise Dice coefficients between sessions (i.e. 1 and 2, 1 and 3, 1 and 4, 2 and 3, 2 and 4, 3 and 4). As described above, the Dice coefficient was calculated as follows:

$$\text{Dice} = 2.V_{overlap}/(V_1 + V_2).$$

Dice coefficients will be described as low (0.00 to 0.19), low-moderate (0.20 to 0.39), moderate (0.40 to 0.59), moderate-high (0.60 to 0.79) or high (0.80 to 1.00).

The LI and Dice coefficient were calculated in two different sets of regions, both derived from the AAL template (Tzourio-Mazoyer et al., 2002). The first set was a broadly defined set of potential language regions (in both hemispheres). In the frontal lobe, the pars opercularis, pars triangularis, and pars orbitalis of the inferior frontal gyrus (IFG) were included. The precentral gyrus was not included, since it includes speech motor areas that were activated in the paradigms with overt responses. In the temporal lobe, the middle temporal gyrus (MTG) was included in its entirety. The ventral part of the superior temporal gyrus (STG) was included, specifically, any voxels within 8 mm of the MTG. The inferior temporal gyrus and fusiform gyrus were included, except for their ventral posterior parts; specifically, any voxels within 8 mm of the cerebellum were excluded. In the parietal lobe, the supramarginal gyrus and angular gyrus were included. In sum, the language ROI included essentially all potential language regions in both hemispheres, but excluded regions associated with motor, auditory and visual processing. The second set of regions comprised all supratentorial structures, that is, the whole brain except for the cerebellum and brainstem (medulla, pons, midbrain). In a supplementary analysis, the LI and Dice coefficient were also calculated separately in frontal and temporal language regions (and homotopic regions). The frontal ROI was defined as the pars opercularis, triangularis and orbitalis of the IFG in either hemisphere, and the temporal ROI was defined as the MTG and the part of the STG that was within 8 mm of the MTG in either hemisphere.

Because lateralization indices and Dice coefficients of similarity were derived from thresholded activation maps, these measures were impacted by the particular threshold chosen. To investigate this dependence on threshold, each contrast was thresholded at 7 different absolute thresholds: $p < 0.1$, $p < 0.05$, $p < 0.01$, $p < 0.005$, $p < 0.001$, $p < 0.0005$, or $p < 0.0001$. Each contrast was also thresholded at 7 different relative thresholds, such that the proportion of activated supratentorial voxels was 10%, 7.5%, 5%, 4%, 3%, 2% or 1%. It has been shown that approaches involving relative thresholds can improve reliability (Knecht et al., 2003; Voyvodic, 2012; Gross and Binder, 2014).

Finally, four different cluster volume cutoffs were applied: none, 500 mm^3, 1000 mm^3, or 2000 mm^3.

3. Results

3.1. Behavioral measures

In the sentence completion paradigm, participants responded correctly to 92.5 ± 4.5% (SD) of trials. They provided incorrect or "I don't know" type responses to 4.9 ± 4.2% of trials, multiple responses (whether correct or not) to 2.1 ± 2.0% of trials, and no responses to 0.5 ± 0.5% of trials. The mean reaction time on correct auditory trials was 4156 ± 193 ms, and the mean reaction time on correct visual trials was 2585 ± 326 ms.

In the picture naming paradigm, participants responded correctly to 95.2 ± 3.0% (SD) of trials. They provided incorrect or "I don't know" type responses to 1.1 ± 0.8% of trials, multiple responses (whether correct or not) to 3.6 ± 2.4% of trials, and no responses to 0.2 ± 0.3% of trials. The mean reaction time on correct trials was 1164 ± 174 ms.

The other two paradigms did not involve behavioral responses, but all participants confirmed after each scan that they had been attentive throughout, and all were able to answer simple questions about the movies and the audiobook stimuli.

3.2. Neuroimaging results

Activation maps were derived for each paradigm, participant, and time point, with a range of voxelwise thresholds and cluster size cutoffs. For illustration, a voxelwise threshold of $p < 0.005$ and a minimum cluster size of 500 mm^3 were used (Fig. 1). None of the paradigms resulted in consistent left-lateralized frontal and temporal activations, which would be expected in most neurologically normal participants. There were substantial differences between the regions activated by the paradigms, some of which were expected given the different language-related and other processes that are implicated (e.g. speech motor in sentence completion and picture naming, but not the other two paradigms). Test-retest reproducibility was not impressive: substantial variability across time points was evident for every paradigm and participant.

Validity and reliability were quantified for each paradigm as a function of voxelwise threshold, whether absolute or relative, cluster size cutoff, and whether the analysis was restricted to language regions of interest, or the whole supratentorial brain (Fig. 2). The LI and Dice coefficients for the frontal and temporal ROIs are also reported separately in Supplementary Fig. 1.

The *sentence completion* paradigm (Fig. 2a) produced reasonably left-lateralized activation maps when analyses were confined to language ROIs. The LI was highest for moderate voxelwise thresholds. More stringent thresholds sometimes resulted in no activated voxels. Dice coefficients were generally low-to-moderate for the analysis parameters that yielded the highest LIs, and neither frontal nor temporal language areas were activated with high sensitivity when thresholds were tightened. When analyses were carried out over the whole brain, reliability was better, but this reflected robust yet bilateral speech motor-related activations.

The *picture naming* paradigm (Fig. 2b) did not yield left-lateralized activation patterns. Activations were dominated by visual object processing and speech motor control, both of which were bilateral. Picture naming was the most reliable of the four paradigms, especially when analyses were performed over the whole brain, yet this reliability reflected robust mapping of sensory and motor processes, rather than language regions.

The *naturalistic comprehension* paradigm (Fig. 2c) resulted in only modestly left-lateralized patterns of activation, which were generally moderately reliable. Left temporal regions were activated even with stringent voxelwise thresholds, but left frontal regions were consistently activated only when voxelwise thresholds were lenient.

The *narrative comprehension* paradigm (Fig. 2d) produced the most strongly left-lateralized language maps of any of the four paradigms. When analyses were restricted to language regions, activations were more lateralized and more reliable (Dice coefficients were generally moderate). There were evident tradeoffs between validity and reliability. At lower voxelwise thresholds, lateralization was weaker, but activation patterns were more reliable, and left frontal and temporal regions were identified with high sensitivity. At higher voxelwise thresholds, lateralization was stronger, but reliability decreased and sensitivity decreased for left frontal regions (sensitivity for left temporal regions generally remained high).

All imaging data will be provided on request from the corresponding author.

4. Discussion

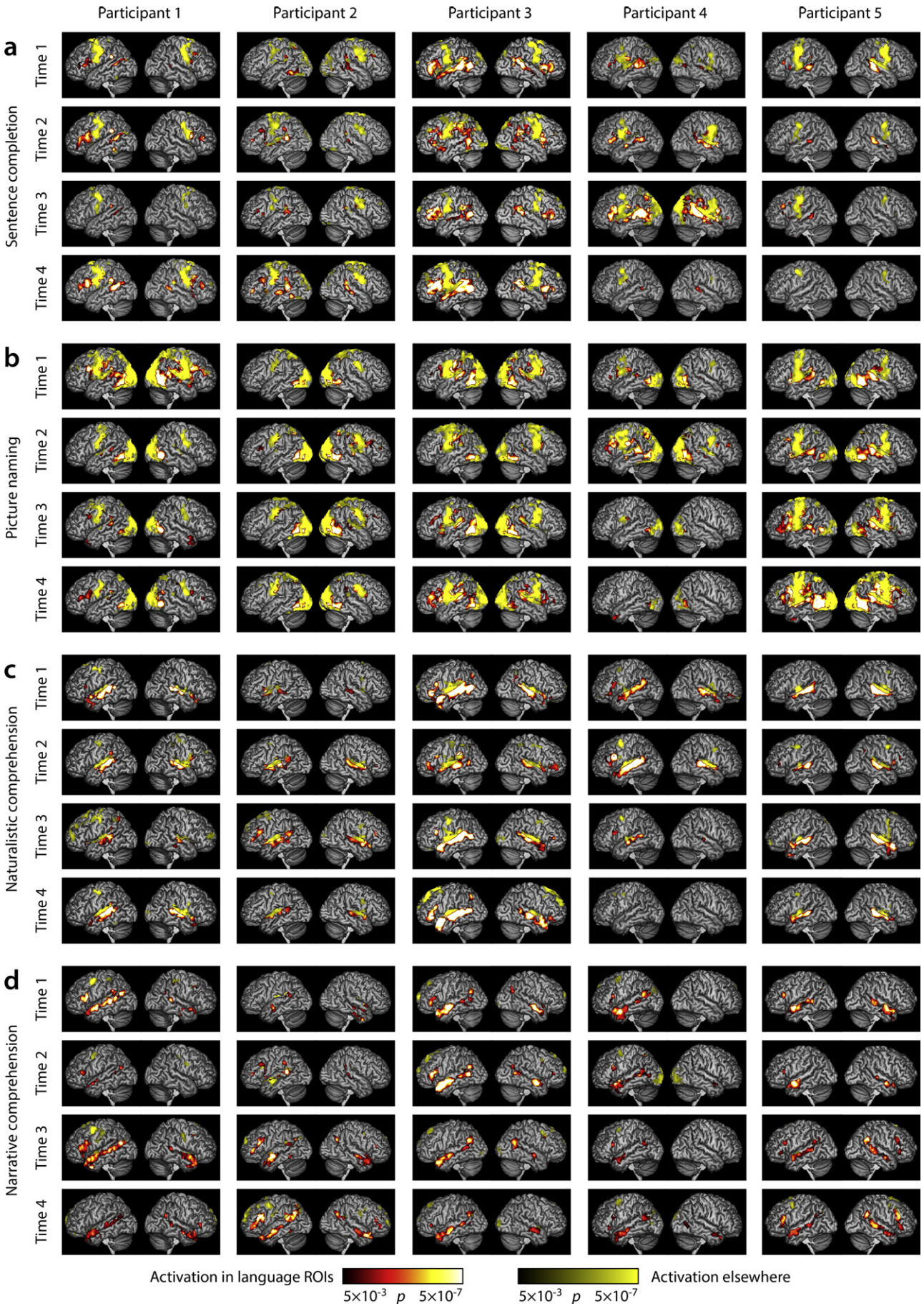
The narrative comprehension and sentence completion paradigms offered the best combination of validity and reliability. However, even for these paradigms, and even with optimal combinations of analysis parameters, there were many scans on which known features of typical language organization were not demonstrated, and test-retest reproducibility was moderate at best.

4.1. Sentence completion paradigm

A plausible set of analysis parameters for the sentence completion paradigm might be a relative voxelwise threshold of 5%, minimum cluster size of 1000 mm^3, and analysis restricted to language regions of interest. Using these parameters, the LI was 0.36, left frontal regions were detected with sensitivity 0.80, left temporal regions were detected with sensitivity 0.95, and the Dice coefficient of similarity was 0.34 (low-moderate). It should be noted that for this paradigm, and the other three paradigms to be discussed below, the actual validity and reliability measures derived from the optimal analysis parameters identified here would need to be evaluated in an independent dataset.

For the sentence completion paradigm, we used overt responses and a conjunction analysis over auditory and visual presentations of incomplete sentences. The requirement for an overt response allows for detailed quantification of performance, which is important in studying individuals with aphasia. However it does result in bilateral speech motor-related activations, which essentially necessitates using an ROI excluding the precentral or postcentral gyri if lateralization is to be demonstrated. This precludes investigation of any involvement of those regions in language processing per se. The conjunction approach implies that the paradigm would be difficult to interpret when auditory and visual comprehension are differentially impaired (which is not common, but certainly attested). Another feasible way to implement a sentence completion paradigm would be to use auditory stimuli only, with an acoustically matched control condition to which participants respond "nothing" or similar.

Previous studies have shown sentence completion paradigms to be moderately effective at activating left-lateralized frontal and temporal language regions. Zacà et al. (2012) studied 41 tumor patients, and reported LIs of 0.23 in a large expressive ROI, and 0.22 in a large receptive ROI, for a covert sentence completion task. Voyvodic (2012) scanned 12 neurologically normal individuals using a covert sentence completion task, and showed much higher LIs (the mean LI was not reported, and depended on analysis parameters) and robust activation of left frontal and left temporal regions. To our knowledge, Dice coefficients have not previously been reported for a sentence completion paradigm. Whalley et al. (2009) investigated test-retest reliability of a sentence completion task, and reported intraclass correlation coefficient (ICC) maps showing maximum reproducibility in frontal and temporal language regions.


Participant 1
Participant 2
Participant 3
Participant 4
Participant 5
a
Sentence completion
b
Picture naming
c
Naturalistic comprehension
d
Narrative comprehension
Time 1
Time 2
Time 3
Time 4
Activation in language ROIs
5×10-3 p 5×10-7
Activation elsewhere
5×10-3 p 5×10-7

4.2. Picture naming paradigm

The picture naming paradigm did not produce left-lateralized activation maps for any combination of parameters. Therefore, it lacks validity as implemented, so it would be meaningless to interpret the reliability measures, which were higher than the other three paradigms.

We did not use any control condition at all, because we were aiming to make the task as simple as possible for patients, and we assumed that the occipito-temporal regions activated by typical control conditions such as processing scrambled objects are sufficiently well established to be readily excluded from analysis.

This may have been a poor decision, since most group studies of picture naming with non-resting baselines do result in somewhat left-lateralized patterns of activation (Price et al., 2005). Explicit investigations of validity of picture naming paradigms have produced mixed results. Some studies have reported good lateralization (Rutten et al., 2002b; Harrington et al., 2006), whereas others have shown poor lateralization (Jansen et al., 2006) and/or frequent lack of activation of frontal or temporal sites (Rau et al., 2007). To our knowledge, Dice coefficients have not previously been reported for a picture naming paradigm, except for in specified regions of interest: Harrington et al. (2006) reported Dice coefficients of ~0.28 in a frontal ROI and ~0.14 in a temporal ROI, and Rau et al. (2007) reported Dice coefficients of 0.48 and 0.49 in the left inferior frontal gyrus pars triangularis and opercularis respectively. Rutten et al. (2002b) reported an idiosyncratic reproducibility statistic for picture naming of up to ~0.27 (depending on threshold); based on the formula used, it appears that the Dice coefficient would be lower. Meltzer et al. (2009) showed high ICCs in many brain regions for picture naming versus rest (for untrained items only), but activation patterns were not lateralized at all, compromising validity. When picture naming was compared to a non-object control condition, activations were more left-lateralized, but ICCs were low in most of the brain.

It may be worthwhile to further investigate the validity and reliability of picture naming paradigms with various control conditions, but findings to date suggest that picture naming is not the most promising language mapping paradigm.

4.3. Naturalistic comprehension paradigm

A plausible set of analysis parameters for the naturalistic comprehension paradigm might be a relative voxelwise threshold of 5%, minimum cluster size of 1000 mm^3, and analysis restricted to language regions of interest. Using these parameters, the LI was 0.20, left frontal regions were detected with sensitivity 0.75, left temporal regions were detected with sensitivity 1.00, and the Dice coefficient of similarity was 0.50 (moderate).

We developed the naturalistic comprehension paradigm in the hope that it might have utility for mapping language regions in severely impaired patients. In our current implementation, the activation patterns were only modestly lateralized, suggesting that the paradigm lacks validity. It appears that although we included a covariate modeling auditory power, bilateral regions involved in auditory processing were activated in addition to language regions. A more sophisticated set of auditory covariates modeling power in different frequency ranges, as well as temporal characteristics, might enable auditory activity to be excluded more effectively. However such covariates might be excessively collinear with the language explanatory variable, because the spectrotemporal characteristics of language differ from most of the auditory content of the television series. Moreover, there may be other differences between segments containing language and those not containing language, for instance, presence or absence of faces, or extent of social interaction.

The naturalistic comprehension paradigm has not previously been used for language mapping, to our knowledge, so no previous studies have investigated the validity or reliability of any similar paradigm.

4.4. Narrative comprehension paradigm

A plausible set of analysis parameters for the naturalistic comprehension paradigm would be a relative voxelwise threshold of 3%, minimum cluster size of 500 mm^3, and analysis restricted to language regions of interest. Using these parameters, the LI was 0.37, left frontal regions were detected with sensitivity 0.90, left temporal regions were detected with sensitivity 1.00, and the Dice coefficient of similarity was 0.41 (moderate).

In this paradigm, we used sparse sampling, which sacrifices some power but avoids the problem that some participants are unable to hear stimuli over scanner noise. We included ten silent intervals (95 s of scan time), that were not used in the analysis in this study. In practice, it is sometimes helpful to be able to compare narrative comprehension to a resting baseline as well as to an acoustically matched baseline, but if that is not necessary, then power could be increased by omitting silent intervals. We used a continuous narrative across the four time points rather than counterbalancing narratives by sessions. This design was chosen based on the idea that a richer narrative context would facilitate activation of anterior temporal language regions (Xu et al., 2005; Binder et al., 2011), but has the disadvantage that narrative segments might not be perfectly matched.

Narrative comprehension tasks with acoustically matched control conditions have been shown to produce quite strongly left-lateralized activation patterns. Harrington et al. (2006) reported LIs of ~0.45 in an inferior frontal ROI and ~0.79 in a temporoparietal ROI. Maldjian et al. (2002) reported a lower LI of 0.19. To our knowledge, Dice coefficients have not previously been reported for a narrative comprehension paradigm, except for in specified regions of interest: Harrington et al. (2006) reported Dice coefficients of ~0.16 and ~0.35 in the inferior frontal and temporoparietal ROIs respectively. Maldjian et al. (2002) also reported a global reproducibility statistic of 0.50, however is not readily interpretable since it was calculated differently from a Dice coefficient. Higher Dice coefficients have been reported for auditory comprehension tasks in two studies that did not have auditory control conditions (Gonzalez-Castillo and Talavage, 2011; Maïza et al., 2011); these contrasts yield bilateral superior temporal activations, which are robust yet reflect primarily auditory processing rather than language.

4.5. Impact of analysis parameters

The analysis parameters we investigated impacted validity and reliability in mostly consistent ways. The various parameter choices also have implications for interpretation depending on the clinical or research context, which need to be considered.

More stringent voxelwise thresholds and larger cluster size cutoffs tended to result in higher LIs, suggesting better validity. This probably reflects appropriate exclusion of spurious activations that would not be expected to be lateralized. However as thresholds became more stringent, Dice coefficients were reduced, sensitivity was reduced for the detection of left frontal and left temporal language regions, and sometimes no voxels were activated at all, leading to undefined LIs. Therefore, in practice, moderate voxelwise thresholds and cluster cutoffs appear to offer the best balance of validity and reliability.

Fig. 1. Activation maps for (a) sentence completion, (b) picture naming, (c) naturalistic comprehension, and (d) narrative comprehension. Each participant is shown in a column, with the four time points arranged from top to bottom. These activation maps were thresholded at voxelwise $p < 0.005$, with a minimum cluster size of 500 mm^3. Activations within the language ROIs are depicted in the hot color scale, while those elsewhere are depicted in yellow. Surface renderings were created with MRIcron (version 0.20140804.1 ~ dfsg.1–1 ~ nd14.04 + 1) with a search depth of 16 voxels.

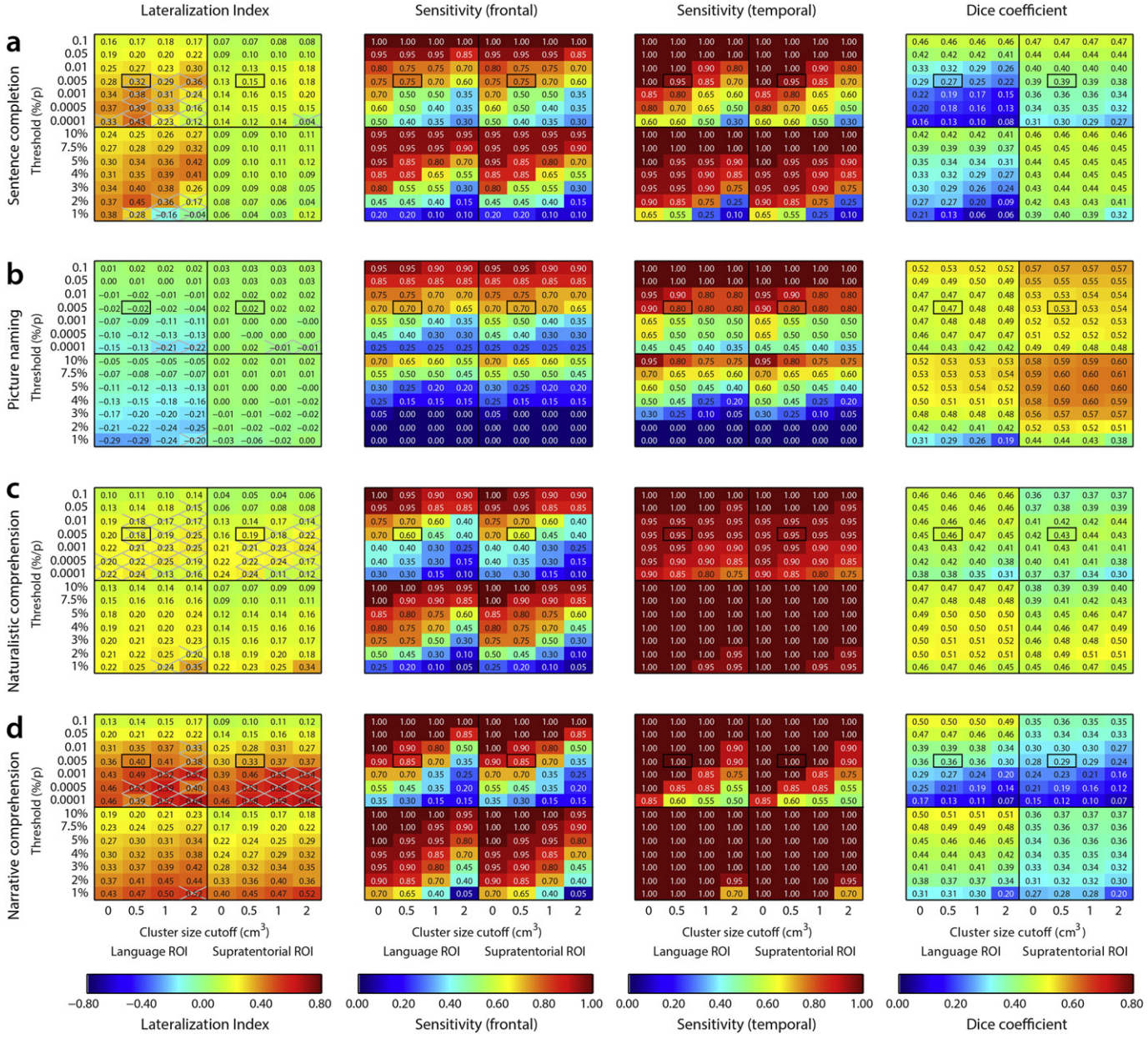


Fig. 2. Validity and reliability of (a) sentence completion, (b) picture naming, (c) naturalistic comprehension, and (d) narrative comprehension. The four columns show laterality indices, proportion of scans with left frontal activation, proportion of scans with left temporal activation, and Dice coefficients of similarity. Within each matrix, absolute and relative threshold choices are shown on the *y* axis, and cluster size cutoffs and regions of interest (language regions only, or whole supratentorial brain) on the *x* axis. The analysis parameters used in Fig. 1 are outlined with black rectangles. Grey crosses indicate that one or more laterality indices were undefined for one or more participants, in which case the reported means exclude those scans.

Relative (percentage of brain) rather than absolute (fixed *p* value) voxelwise thresholds tended to significantly improve reliability, without having any negative impact on validity. Several previous studies have reported similar findings (e.g. Knecht et al., 2003; Voyvodic, 2012; Gross and Binder, 2014) It is not surprising that relative thresholds improve reliability, because by holding the total amount of activation constant, they remove a substantial source of test-retest variability. The total amount of activation reflects in large part factors such as alertness, attention, caffeine consumption, degree of head motion, and so on, in other words, not genuine differences in the extent of language processing regions. In pre-surgical applications, relative thresholding can be recommended, because the primary goal is to localize language regions once and once only, and their precise extent is of secondary concern, and cannot be determined reliably anyway due to the generic factors mentioned above. On the other hand, in research contexts such as investigations of language reorganization after stroke, relative thresholding has a potential downside, because it is plausible that over the course of recovery there could be real changes in the extent of regions devoted to language processing. Such changes would not be detectable when using a method that fixes the extent of language regions in advance. One possible way around this could be to add a highly robust sensory or motor control task to each session such as viewing a checkerboard or finger tapping. Activation to this control task could be used to derive a measure of functional sensitivity for the session, which could be used to adjust language maps instead of fixing their extent in advance.

When analyses were restricted to a priori likely language regions, both validity and reliability were generally better. This is not surprising, since many activations outside of likely language regions would be spurious. Moreover, some paradigms recruited bilateral sensorimotor

regions, notably sentence completion, due to the overt spoken response, and picture naming, due to the overt spoken response and visual object processing. For these types of paradigms, LIs were greatly reduced unless non-language regions were excluded. However, restricting analyses to a priori likely language regions could be a risky strategy in clinical and research contexts, since functional reorganization due to neurological conditions may result in brain areas outside of the typical language network becoming involved in language processing. Therefore, the positive impact on validity and reliability of restricting analyses to language regions needs to be balanced against this significant limitation.

4.6. General discussion

The two paradigms which offered the best combination of validity and reliability were sentence completion and narrative comprehension. However even under optimal analysis parameters, activation patterns were only moderately left-lateralized (LI = 0.36 for sentence completion, 0.37 for narrative comprehension), left frontal and left temporal language regions were not always detected, and Dice coefficients were only low-moderate to moderate (0.34 for sentence completion, 0.41 for narrative comprehension). To put these numbers in perspective, a voxel activated in a language comprehension paradigm has a 41% chance of being activated if the same individual is scanned on another occasion. This is not really a high level of confidence. Moreover, these results were obtained using relative voxelwise thresholds and ROIs restricted to broad a priori language regions; both of which approaches entail limitations for investigation of language reorganization over time, as discussed above.

There is a clear need for similar investigation of different paradigms that may have better psychometric properties. Paradigms involving semantic judgments appear to be particularly promising, as they have been shown to be strongly lateralizing (Binder et al., 1996; Binder et al., 2008; Szaflarski et al., 2008; Fesl et al., 2010; Janecek et al., 2013b), and have good test-retest reproducibility. The only two studies to our knowledge that have reported test-retest Dice coefficients of similarity for well controlled language tasks across the whole language network or the whole brain both used semantic tasks. Fernández et al. (2003) examined test-retest reliability in 12 patients with drug-resistant focal epilepsy; the two scanning sessions were in each case performed on the same day. The language task was a semantic matching task and the control task was a perceptual matching task. The mean Dice coefficient ranged from 0.43 to 0.49 depending on the voxelwise threshold. Fesl et al. (2010) investigated 39 healthy participants who were evenly divided into three groups of right-handed, bimanual and left-handed. Two fMRI sessions were carried out about one week apart. A semantic task was compared to a perceptually matched control task. The mean Dice coefficient in a large ROI comprising a very broad network of potential language regions and right hemisphere homotopic regions was 0.61. While semantic judgment tasks therefore appear to have good validity and reliability, a major concern is whether they can be performed by individuals with aphasia. An important avenue for future research will be investigating the abilities of people with aphasia to perform semantic judgment tasks, and potentially adapting these tasks for this population.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.nicl.2016.03.015.

Acknowledgments

We thank Scott Squire, Angelica McCarron and Sarah Olson for technical assistance, Andrew DeMarco for helpful discussions, three anonymous reviewers for their constructive comments, and the individuals who participated in the study. This work was supported by the National Institute on Deafness and Other Communication Disorders (NIH R01 DC013270).

References

Abel, S., Weiller, C., Huber, W., Willmes, K., Specht, K., 2015. Therapy-induced brain reorganization patterns in aphasia. Brain 138, 1097–1112.
Ashburner, J., Friston, K.J., 2005. Unified segmentation. NeuroImage 26, 839–851.
Bauer, J., 2004. Hope Was Here [Compact Disc] Lamia, J., reader Random House/Listening Library, New York.
Bauer, P.R., Reitsma, J.B., Houweling, B.M., Ferrier, C.H., Ramsey, N.F., 2014. Can fMRI safely replace the Wada test for preoperative assessment of language lateralisation? A meta-analysis and systematic review. J. Neurol. Neurosurg. Psychiatr. 85, 581–588.
Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans. Med. Imaging 23, 137–152.
Benke, T., Köylü, B., Visani, P., Karner, E., Brenneis, C., Bartha, L., Trinka, E., Trieb, T., Felber, S., Bauer, G., Chemelli, A., Willmes, K., 2006. Language lateralization in temporal lobe epilepsy: a comparison between fMRI and the Wada Test. Epilepsia 47, 1308–1319.
Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. 1191, 133–155.
Berl, M.M., Zimmaro, L.A., Khan, O.I., Dustin, I., Ritzl, E., Duke, E.S., Sepeta, L.N., Sato, S., Theodore, W.H., Gaillard, W.D., 2014. Characterization of atypical language activation patterns in focal epilepsy. Ann. Neurol. 75, 33–42.
Billingsley-Marshall, R., Simos, P., Papanicolaou, A., 2004. Reliability and validity of functional neuroimaging techniques for identifying language-critical areas in children and adults. Dev. Neuropsychol. 26, 541–563.
Binder, J.R., Swanson, S.J., Hammeke, T.A., Morris, G.L., Mueller, W.M., Fischer, M., Benbadis, S., Frost, J.A., Rao, S.M., Haughton, V.M., 1996. Determination of language dominance using functional MRI: a comparison with the Wada test. Neurology 46, 978–984.
Binder, J.R., Swanson, S.J., Hammeke, T.A., Sabsevitz, D.S., 2008. A comparison of five fMRI protocols for mapping speech comprehension systems. Epilepsia 49, 1980–1997.
Binder, J.R., Gross, W.L., Allendorfer, J.B., Bonilha, L., Chapin, J., Edwards, J.C., Grabowski, T.J., Langfitt, J.T., Loring, D.W., Lowe, M.J., Koenig, K., Morgan, P.S., Ojemann, J.G., Rorden, C., Szaflarski, J.P., Tivarus, M.E., Weaver, K.E., 2011. Mapping anterior temporal lobe language areas with fMRI: a multicenter normative study. NeuroImage 54, 1465–1475.
Bizzi, A., Blasi, V., Falini, A., Ferroli, P., Cadioli, M., Danesi, U., Aquino, D., Marras, C., Caldiroli, D., Broggi, G., 2008. Presurgical functional MR imaging of language and motor functions: validation with intraoperative electrocortical mapping. Radiology 248, 579–589.
Block, C.K., Baldwin, C.L., 2010. Cloze probability and completion norms for 498 sentences: behavioral and neural validation using event-related potentials. Behav. Res. Methods 42, 665–670.
Brainard, D.H., 1997. The psychophysics toolbox. Spat. Vis. 10, 433–436.
Brannen, J.H., Badie, B., Moritz, C.H., Quigley, M., Meyerand, M.E., Haughton, V.M., 2001. Reliability of functional MR imaging with word-generation tasks for mapping Broca's area. AJNR Am. J. Neuroradiol. 22, 1711–1718.
Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173.
Crinion, J., Price, C.J., 2005. Right anterior superior temporal activation predicts auditory sentence comprehension following aphasic stroke. Brain 128, 2858–2871.
Eaton, K.P., Szaflarski, J.P., Altaye, M., Ball, A.L., Kissela, B.M., Banks, C., Holland, S.K., 2008. Reliability of fMRI for studies of language in post-stroke aphasia subjects. NeuroImage 41, 311–322.
Fernández, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., Klaver, P., Ruhlmann, J., Reul, J., Elger, C.E., 2003. Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. Neurology 60, 969–975.
Fesl, G., Bruhns, P., Rau, S., Wiesmann, M., Ilmberger, J., Kegel, G., Brueckmann, H., 2010. Sensitivity and reliability of language laterality assessment with a free reversed association task—a fMRI study. Eur. Radiol. 20, 683–695.
Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12, 189–198.
Freaks and Geeks 1999–2000, Television series. Apatow Productions and DreamWorks Television, Los Angeles, California. Created by Paul Feig.
Fridriksson, J., Baker, J.M., Moser, D., 2009. Cortical mapping of naming errors in aphasia. Hum. Brain Mapp. 30, 2487–2498.
Giussani, C., Roux, F.-E., Ojemann, J., Sganzerla, E.P., Pirillo, D., Papagno, C., 2010. Is preoperative functional magnetic resonance imaging reliable for language areas mapping in brain tumor surgery? Review of language functional magnetic resonance imaging and direct cortical stimulation correlation studies. Neurosurgery 66, 113–120.
Gonzalez-Castillo, J., Talavage, T.M., 2011. Reproducibility of fMRI activations associated with auditory sentence comprehension. NeuroImage 54, 2138–2155.
Gross, W.L., Binder, J.R., 2014. Alternative thresholding methods for fMRI data optimized for surgical planning. NeuroImage 84, 554–561.
Harrington, G.S., Buonocore, M.H., Farias, S.T., 2006. Intrasubject reproducibility of functional MR imaging activation in language tasks. AJNR Am. J. Neuroradiol. 27, 938–944.
Janecek, J.K., Swanson, S.J., Sabsevitz, D.S., Hammeke, T.A., Raghavan, M., Mueller, W., Binder, J.R., 2013a. Naming outcome prediction in patients with discordant Wada and fMRI language lateralization. Epilepsy Behav. 27, 399–403.
Janecek, J.K., Swanson, S.J., Sabsevitz, D.S., Hammeke, T.A., Raghavan, M., E. Rozman, M., Binder, J.R., 2013b. Language lateralization by fMRI and Wada testing in 229 patients with epilepsy: rates and predictors of discordance. Epilepsia 54, 314–322.
Jansen, A., Menke, R., Sommer, J., Förster, A.F., Bruchmann, S., Hempleman, J., Weber, B., Knecht, S., 2006. The assessment of hemispheric lateralization in functional MRI—robustness and reproducibility. NeuroImage 33, 204–217.

Kelly Jr., R.E., Alexopoulos, G.S., Wang, Z., Gunning, F.M., Murphy, C.F., Morimoto, S.S., Kanellopoulos, D., Jia, Z., Lim, K.O., Hoptman, M.J., 2010. Visual inspection of independent components: defining a procedure for artifact removal from fMRI data. J. Neurosci. Methods 189, 233–245.

Kho, K.H., Leijten, F.S.S., Rutten, G.-J., Vermeulen, J., Van Rijen, P., Ramsey, N.F., 2005. Discrepant findings for Wada test and functional magnetic resonance imaging with regard to language function: use of electrocortical stimulation mapping to confirm results. Case report. J. Neurosurg. 102, 169–173.

Kiran, S., Ansaldo, A., Bastiaanse, R., Cherney, L.R., Howard, D., Faroqi-Shah, Y., Meinzer, M., Thompson, C.K., 2013. Neuroimaging in aphasia treatment research: standards for establishing the effects of treatment. NeuroImage 76, 428–435.

Knecht, S., Jansen, A., Frank, A., van Randenborgh, J., Sommer, J., Kanowski, M., Heinze, H.J., 2003. How atypical is atypical language dominance? NeuroImage 18, 917–927.

Lanzenberger, R., Wiest, G., Geissler, A., Barth, M., Ringl, H., Wöber, C., Gartus, A., Baumgartner, C., Beisteiner, R., 2005. FMRI reveals functional cortex in a case of inconclusive Wada testing. Clin. Neurol. Neurosurg. 107, 147–151.

Lund, K., Burgess, C., 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Meth. Ins. C. 28, 203–208.

Maïza, O., Mazoyer, B., Hervé, P.-Y., Razafimandimby, A., Dollfus, S., Tzourio-Mazoyer, N., 2011. Reproducibility of fMRI activations during a story listening task in patients with schizophrenia. Schizophr. Res. 128, 98–101.

Maldjian, J.A., Laurienti, P.J., Driskill, L., Burdette, J.H., 2002. Multiple reproducibility indices for evaluation of cognitive functional MR imaging paradigms. AJNR Am. J. Neuroradiol. 23, 1030–1037.

Meinzer, M., Beeson, P.M., Cappa, S., Crinion, J., Kiran, S., Saur, D., Parrish, T., Crosson, B., Thompson, C.K., Neuroimaging in Aphasia Treatment Research Workshop, 2013. Neuroimaging in aphasia treatment research: consensus and practical guidelines for data analysis. NeuroImage 73, 215–224.

Meltzer, J.A., Postman-Caucheteux, W.A., McArdle, J.J., Braun, A.R., 2009. Strategies for longitudinal neuroimaging studies of overt language production. NeuroImage 47, 745–755.

Partovi, S., Konrad, F., Karimi, S., Rengier, F., Lyo, J.K., Zipp, L., Nennig, E., Stippich, C., 2012. Effects of covert and overt paradigms in clinical language fMRI. Acad. Radiol. 19, 518–525.

Pelli, D.G., 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis. 10, 437–442.

Penfield, W., Roberts, L., 1959. Speech and Brain-Mechanisms. Princeton University Press, Princeton, NJ.

Postman-Caucheteux, W.A., Birn, R.M., Pursley, R.H., Butman, J.A., Solomon, J.M., Picchioni, D., McArdle, J., Braun, A.R., 2010. Single-trial fMRI shows contralesional activity linked to overt naming errors in chronic aphasic patients. J. Cogn. Neurosci. 22, 1299–1318.

Pouratian, N., Bookheimer, S.Y., Rex, D.E., Martin, N.A., Toga, A.W., 2002. Utility of preoperative functional magnetic resonance imaging for identifying language cortices in patients with vascular malformations. J. Neurosurg. 97, 21–32.

Price, C.J., Devlin, J.T., Moore, C.J., Morton, C., Laird, A.R., 2005. Meta-analyses of object naming: effect of baseline. Hum. Brain Mapp. 25, 70–82.

Price, C.J., Crinion, J., Friston, K.J., 2006. Design and analysis of fMRI studies with neurologically impaired patients. J. Magn. Reson. Imaging 23, 816–826.

Rau, S., Fesl, G., Bruhns, P., Havel, P., Braun, B., Tonn, J.-C., Ilmberger, J., 2007. Reproducibility of activations in Broca area with two language tasks: a functional MR imaging study. AJNR Am. J. Neuroradiol. 28, 1346–1353.

Rombouts, S.A., Barkhof, F., Hoogenraad, F.G., Sprenger, M., Valk, J., Scheltens, P., 1997. Test-retest analysis with functional MR of the activated area in the human visual cortex. AJNR Am. J. Neuroradiol. 18, 1317–1322.

Rossion, B., Pourtois, G., 2004. Revisiting Snodgrass and Vanderwart's object pictorial set: the role of surface detail in basic-level object recognition. Perception 33, 217–236.

Rutten, G.J.M., Ramsey, N.F., van Rijen, P.C., Noordmans, H.J., van Veelen, C.W.M., 2002a. Development of a functional magnetic resonance imaging protocol for intraoperative localization of critical temporoparietal language areas. Ann. Neurol. 51, 350–360.

Rutten, G.J.M., Ramsey, N.F., van Rijen, P.C., van Veelen, C.W.M., 2002b. Reproducibility of fMRI-determined language lateralization in individual subjects. Brain Lang. 80, 421–437.

Sanai, N., Mirzadeh, Z., Berger, M.S., 2008. Functional outcome after language mapping for glioma resection. N. Engl. J. Med. 358, 18–27.

Seghier, M.L., Lazeyras, F., Pegna, A.J., Annoni, J.M., Zimine, I., Mayer, E., Michel, C.M., Khateb, A., 2004. Variability of fMRI activation during a phonological and semantic language task in healthy subjects. Hum. Brain Mapp. 23, 140–155.

Seghier, M.L., Kherif, F., Josse, G., Price, C.J., 2011. Regional and hemispheric determinants of language laterality: implications for preoperative fMRI. Hum. Brain Mapp. 32, 1602–1614.

Snodgrass, J.G., Vanderwart, M., 1980. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. J. Exp. Psychol. 6, 174–215.

Szaflarski, J.P., Holland, S.K., Jacola, L.M., Lindsell, C., Privitera, M.D., Szaflarski, M., 2008. Comprehensive presurgical functional MRI language evaluation in adult patients with epilepsy. Epilepsy Behav. 12, 74–83.

Thompson, C.K., den Ouden, D.-B., 2008. Neuroimaging and recovery of language in aphasia. Curr. Neurol. Neurosci. Rep. 8, 475–483.

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

Tzourio-Mazoyer, N., Petit, L., Razafimandimby, A., Crivello, F., Zago, L., Jobard, G., Joliot, M., Mellet, E., Mazoyer, B., 2010. Left hemisphere lateralization for language in right-handers is controlled in part by familial sinistrality, manual preference strength, and head size. J. Neurosci. 30, 13314–13318.

Voyvodic, J.T., 2012. Reproducibility of single-subject fMRI language mapping with AMPLE normalization. J. Magn. Reson. Imaging 36, 569–580.

Whalley, H.C., Gountouna, V.-E., Hall, J., McIntosh, A.M., Simonotto, E., Job, D.E., Owens, D.G.C., Johnstone, E.C., Lawrie, S.M., 2009. fMRI changes over time and reproducibility in unmedicated subjects at high genetic risk of schizophrenia. Psychol. Med. 39, 1189.

Wilson, S.M., Isenberg, A.L., Hickok, G., 2009. Neural correlates of word production stages delineated by parametric modulation of psycholinguistic variables. Hum. Brain Mapp. 30, 3596–3608.

Wilson, S.M., Lam, D., Babiak, M.C., Perry, D.W., Shih, T., Hess, C.P., Berger, M.S., Chang, E.F., 2015. Transient aphasias after left hemisphere resective surgery. J. Neurosurg. 123, 581–593.

Woermann, F.G., Jokeit, H., Luerding, R., Freitag, H., Schulz, R., Guertler, S., Okujava, M., Wolf, P., Tuxhorn, I., Ebner, A., 2003. Language lateralization by Wada test and fMRI in 100 patients with epilepsy. Neurology 61, 699–701.

Worsley, K.J., Liao, C.H., Aston, J., Petre, V., Duncan, G.H., Morales, F., Evans, A.C., 2002. A general statistical analysis for fMRI data. NeuroImage 15, 1–15.

Xu, J., Kemeny, S., Park, G., Frattali, C., Braun, A., 2005. Language in context: emergent features of word, sentence, and narrative comprehension. NeuroImage 25, 1002–1015.

Yetkin, F.Z., Mueller, W.M., Morris, G.L., McAuliffe, T.L., Ulmer, J.L., Cox, R.W., Daniels, D.L., Haughton, V.M., 1997. Functional MR activation correlated with intraoperative cortical mapping. AJNR Am. J. Neuroradiol. 18, 1311–1315.

Zacà, D., Nickerson, J.P., Deib, G., Pillai, J.J., 2012. Effectiveness of four different clinical fMRI paradigms for preoperative regional determination of language lateralization in patients with brain tumors. Neuroradiology 54, 1015–1025.