

Efficient functional localization of language regions in the brain

Jayden J. Lee^a, Terri L. Scott^b, Tyler K. Perrachione^{a,*}

^a Department of Speech, Language, and Hearing Sciences, Boston University, 635 Commonwealth Ave., Boston, MA 02215, United States

^b Department of Neurological Surgery, University of California – San Francisco, San Francisco, CA, United States

ARTICLE INFO

Keywords:

Language
Functional localizer
Functional selectivity
Test-retest reliability
Single-subject design

ABSTRACT

Important recent advances in the cognitive neuroscience of language have been made using functional localizers to demarcate language-selective regions in individual brains. Although single-subject localizers offer insights that are unavailable in classic group analyses, they require additional scan time that imposes costs on investigators and participants. In particular, the unique practical challenges of scanning children and other special populations has led to less adoption of localizers for neuroimaging research with these theoretically and clinically important groups. Here, we examined how measurements of the spatial extent and functional response profiles of language regions are affected by the duration of an auditory language localizer. We compared how parametrically smaller amounts of data collected from one scanning session affected (i) consistency of group-level whole-brain parcellations, (ii) functional selectivity of subject-level activation in individually defined functional regions of interest (fROIs), (iii) sensitivity and specificity of subject-level whole-brain and fROI activation, and (iv) test-retest reliability of subject-level whole-brain and fROI activation. For many of these metrics, the localizer duration could be reduced by 50-75% while preserving the stability and reliability of both the spatial extent and functional response profiles of language areas. These results indicate that, for most measures relevant to cognitive neuroimaging studies, the brain's language network can be localized just as effectively with 3.5 min of scan time as it can with 12 min. Minimizing the time required to reliably localize the brain's language network allows more effective localizer use in situations where each minute of scan time is particularly precious.

1. Introduction

A longstanding question in neuroscience is the extent to which cognitive faculties can be selectively attributed to distinct brain regions and vice-versa, with the left-hemisphere organization of language representing one of the earliest (Broca, 1861) and most widely studied (Hickok & Poeppel, 2007; Price, 2012) aspects of the functional organization of the brain. One particularly effective neuroimaging approach to investigating functional neuroanatomy in individual subjects has been through the use of *functional localizers* (Saxe et al., 2006). The basic premise of a functional localizer is that functional regions of interest (fROIs) can be specified in the brains of individuals based on subject-specific patterns of response to a stimulus or task of interest. The response properties of those fROIs can then be investigated across other stimuli and tasks to determine their profiles of selective vs. shared responses. Functional localizers have played an important role in understanding the functional organization and selectivity of cortical regions that are responsible for numerous perceptual and cognitive domains,

such as the perception of faces (Fox et al., 2009; Kanwisher et al., 1997), bodies (Downing et al., 2001; Ross et al., 2020), voices (Belin et al., 2000; Pernet et al., 2015), and printed words (Cohen et al., 2002; Dehaene et al., 2002), as well as working memory (Somers et al., 2021) and core linguistic processing (Fedorenko et al., 2011; Fedorenko et al., 2010).

A unique contribution of functional localizers versus other analytical approaches is that researchers can define and examine regions of interest in the brains of individuals based on their patterns of functional response rather than their macroanatomical (e.g., gyrus or sulcus) or stereotactic location. In this way, functional localization helps overcome the substantial variability in structure-function correspondence across individual brains (Brett et al., 2002), which is obfuscated by the traditional group analyses that are based on strict stereotactic voxelwise correspondence across brains (Fedorenko, 2021). In particular, functional localizers offer an encouraging new avenue for understanding the functional architecture of the cortical language network, where traditional group analyses have led to conflicting conclusions about the

* Corresponding author.

E-mail address: tkp@bu.edu (T.K. Perrachione).

functional specialization or selectivity of classic language areas (Fedorenko et al., 2011; Nieto-Castañón & Fedorenko, 2012).

The most widely implemented functional language localizer involves an in-scanner task that contrasts brain activation elicited by reading meaningful sentences vs. sequences of pronounceable but meaningless nonwords in order to identify the brain regions sensitive to word- and sentence-level meaning (Fedorenko et al., 2010). This localizer contrast effectively dissociates brain areas that are selectively responsive to linguistic processing from those that are responsive to other tasks involving domain-general cognitive functions such as working memory and executive control, which are situated in immediately adjacent frontal, temporal, and parietal cortices (Blank et al., 2014; Fedorenko et al., 2012). For example, this language localizer has shed new light on classical conundrums such as whether language and thought are subserved by distinct brain regions (Fedorenko & Varley, 2016), whether syntactic and semantic processing involve distinct brain regions (Fedorenko et al., 2020), and whether language comprehension recruits neural tissue that supports other domain-general cognitive faculties (Diachek et al., 2020; Mineroff et al., 2018).

However, a task that involves reading printed words and pseudo-words is not ideal when considering many special populations, including young children who have not yet learned to read, children and adults who have difficulty with reading or with decoding nonwords (i.e., those with dyslexia; Gabrieli, 2009), and individuals who have difficulties with sustained attention to tasks. To address these limitations, an auditory version of the language localizer has been developed that contrasts participants' neural activation while passively listening to brief recordings of meaningful speech vs. incomprehensible degraded speech (Scott et al., 2017). The spatial extent and functional response profiles of language regions identified by this auditory language localizer are highly consistent with those obtained using the original reading task. However, while an auditory localizer resolves the extraneous literacy and cognitive demands of its predecessor, its prescribed duration of nearly 13 min remains onerous in the context of neuroimaging research with pediatric and special populations. In these populations, achieving comfort and compliance with the in-scanner environment can be challenging (Fassbender et al., 2017; Greene et al., 2016), making every minute of active scan time especially precious (Meissner et al., 2020). In addition to these practical challenges, there are substantial monetary costs involved in operating an MRI scanner. Such expense may lead researchers to squeeze their preferred experimental tasks into every available minute of scan time, rather than setting aside some of these valuable minutes for a localizer (Friston et al., 2006). Without a localizer, researchers must look for group-level effects using unconstrained whole-brain or anatomical ROI-based analyses, in spite of the well-known inferential limitations of these techniques (Fedorenko, 2021; Saxe et al., 2006).

In the real-world conduct of neuroimaging research where the availability of scan time is practically and financially limited, the appeal of functional localizers varies inversely with the time needed to obtain these scans. Correspondingly, any improvement in localizer efficiency confers both economic *and* theoretical benefits to researchers: For a researcher planning to recruit 30 participants for a new fMRI study today, reducing a localizer's scan time from 12 min to 4 min can save nearly \$2500 in scanner operating expenses over the course of the study. This recovered value could instead be applied to obtaining additional, theoretically-motivated neuroimaging data from other tasks (Friston et al., 2006), scanning additional participants to improve statistical power (Mumford, 2012), or paying increasingly exorbitant open-access publishing fees (Manca et al., 2017). However, the utility of a functional localizer depends on its sensitivity, specificity, and reliability (Berman et al., 2010; Elliott et al., 2021), and given the inherent noisiness of the fMRI signal, collecting more data is perhaps the single most effective approach to improving within-subject reliability (Nee, 2019; Perrachione & Ghosh, 2013). This is particularly relevant for functional localizers, which obviate many of the other sources of between-subjects

variability in fMRI, particularly those due to spatial variability in functional neuroanatomy. Thus, optimizing the amount of scan time to devote to collecting a functional localizer reflects a complex balance of numerous scientific and practical demands.

In this study, we sought to characterize how reducing the scan time of one widely-used auditory language localizer (Scott et al., 2017) affects its measurement of the spatial extent and functional response of language-related brain areas. We characterized changes in individual subjects' activation measured from parametrically smaller amounts of data obtained during one scanning session (i.e., shorter localizer scan times). A limitation of this approach is that, in many cases, it precludes inferential statistical comparisons of samples of data that partially overlap. However, we chose to analyze increasingly shorter subsets of a single localizer stimulation paradigm, rather than compare several different localizer paradigms of different durations, because we believe this approach parallels how researchers carry out real-world design evaluation when they are weighing the opportunity costs of conducting longer vs. shorter in-scanner tasks.

Specifically, we compared increasingly shorter amounts of data sub-selected from an original localizer acquisition vs. its full duration with respect to (i) the consistency of group-level whole-brain areal parcelations (Julian et al., 2012), (ii) the degree of language selectivity within single-subject fROIs (Fedorenko et al., 2011; 2012), (iii) the sensitivity and specificity of whole-brain activation and fROI location at the single-subject level (Nieto-Castañón & Fedorenko, 2012), and (iv) the test-retest reliability of subject-level whole-brain and fROI functional responses (Gorgolewski et al., 2013). We found that most of these measurements were extremely robust in the face of shorter localizer durations (i.e., the unbiased response selectivity of fROIs for linguistic stimuli, the low likelihood of finding false-positive activation, and the test-retest reliability of the spatial extent and response profile of subjects' language-selective responses between runs), while other measurements were monotonically affected by localizer duration (principally, detection of all true-positive language-selective voxels). By characterizing the effects of localizer duration on the reliability and stability of language-selective response across the whole brain, we reveal the highly selective and highly reliable functional response profile of language regions measured in individual brains. We hope these findings are useful in guiding other researchers to make principled choices regarding the amount of functional language localizer scan time that is optimal to their unique research needs.

2. Material and methods

2.1. Participants

Twenty-four adults (13 female, 11 male; age 19–32; $M = 23.5$ years) completed this study. All participants spoke fluent English and, by self-report, had no history of speech, language, hearing, or neurological disorder or any reading, cognitive, or motor developmental difficulties. All participants gave informed, written consent, and this study was approved and overseen by the Institutional Review Board at Boston University and the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology. Participants received monetary compensation for their participation.

2.2. In-scanner tasks

Each participant performed two tasks as part of this study: a passive listening task to localize the language network (Scott et al., 2017) and a visual-spatial working memory task, which served as a non-linguistic control to examine the language selectivity of the regions identified by the auditory language localizer (Fedorenko et al. 2011). All participants who completed the language localizer and the working memory task also performed a number of additional tasks for separate studies, which are not considered here (Scott & Perrachione, 2019; Scott, 2020).

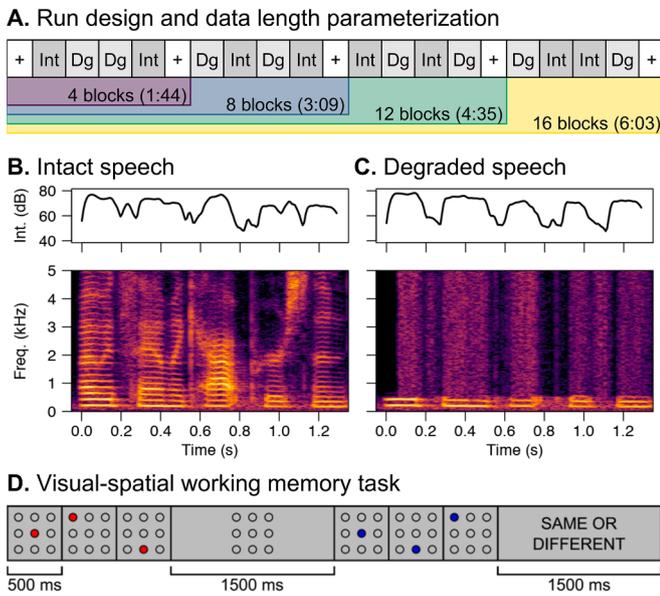


Fig. 1. Language localizer and task design. (A) Each run of the localizer was subdivided into four spans of decreasing amounts of data (Run 1 is shown). The task consisted of passive listening during blocks of intact (Int) and degraded (Dg) speech, as well as resting baseline fixation (+). (B) A brief sample of the auditory stimuli from the Intact speech condition, showing the dynamics of the intensity envelope and spectrotemporal features (spectrogram) of the phrase, “I would say I’m a cat person,” comprising approximately 1.3 s of an 18 s intact speech block. (C) The corresponding stimulus from the Degraded speech condition, in which the speech stimulus was low-pass filtered and white noise was added with a temporally-modulated envelope matching that of the intact speech. (A participant heard either the intact or degraded version of a stimulus derived from a particular source recording, thereby preserving the total incomprehensibility of degraded speech stimuli.) (D) An example 3-item (low-load) visual-spatial working memory trial is shown; the correct response is “different.”

2.3. Language localizer

Participants passively listened to engaging excerpts of natural speech from various long-form interviews and podcasts (e.g., TED Talks, The Moth Podcast) and acoustically degraded versions of similar excerpts in a blocked design (Scott et al., 2017). From each recording of natural speech (Fig. 1B), a corresponding degraded version was created (Fig. 1C) that effaced all phonological and linguistic content while preserving the intensity of time-varying acoustic stimulation (Stoppelman et al., 2013). Participants only heard either the intact or degraded version of any particular audio clip. Degraded speech stimuli were low-pass filtered copies of the intact stimuli (using a pass-band frequency cutoff of 500 Hz) to which we added broadband white noise that had been multiplied by the amplitude envelope of the corresponding intact recording to produce matched time-varying fluctuations of acoustic intensity. The noise track was then low-pass filtered to ‘soften’ the highest frequencies using a pass-band frequency of 8,000 Hz and a stop frequency of 10,000 Hz and added to the low-pass filtered copies of the clips to construct degraded versions of each intact speech recording. The experience of stimuli in the degraded speech condition is like listening to garbled, unintelligible radio transmissions. All materials for this spoken language localizer, including scripts and transcriptions of the speech excerpts, are available online.¹

One 18 s recording was presented per block of intact or degraded speech. The duration of the full language localizer was 12:18 (min:sec), with each of the two runs lasting 6:08 and consisting of 16 stimulus

blocks (8 intact speech and 8 degraded) and 5 fixation blocks (14 s rest blocks in which participants heard no stimuli).

2.4. Visual-spatial working memory

We also wanted to determine whether and how the language-selective responses of brain areas identified by the localizer were affected by localizer duration. Extensive research has shown that the cortical networks for language and domain-general (multiple demand) cognitive tasks like spatial working memory appear to be strictly dissociated (Diachek et al., 2020; Mineroff et al., 2018). To this end, we also measured participants’ brain responses to a spatial working memory task (derived from the classic Corsi block-tapping test). In this task, participants recognized sequences of dots presented one at a time in a 3 × 3 grid (Scott, 2020) under two levels of visual-spatial working memory load. On each trial of the low-load (3-item sequences) and high-load (6-item sequences) conditions, participants were presented with two sequences of dots and asked to indicate whether the second sequence was identical to the first. Each dot in the first sequence was illuminated in red for 500 ms, followed by a 1.5 s retention interval presentation of the empty grid, and then each dot in the second sequence was illuminated in blue for 500 ms (Fig. 1D). Participants responded to this two-alternative forced choice task by button press when they saw the words “SAME OR DIFFERENT” presented on screen. The *high > low* working memory-load contrast is intended to identify regions that are responsive to increased demands on visual-spatial working memory, which activates the *multiple demand* (MD) network (Blank et al., 2014; Diachek et al., 2020; Mineroff et al., 2018).

Condition order was counterbalanced across the two runs. Each condition included five blocks (4 trials per block) and five, 15 s blocks of rest. Each low-load block lasted 24 s and each high-load block lasted 36 s. Each run lasted 6:18, and all participants completed two runs of this task.

2.5. MRI data acquisition

Structural and functional data were acquired on a whole-body Siemens Trio 3T scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. Participants were situated in a head-first supine position in the scanner. Auditory stimuli were presented over Sennheiser HD 202 MRI-compatible Model S14 headphones and visual cues were presented over an in-scanner projector screen. High-resolution structural images, including a T1-weighted magnetization-prepared rapid gradient-echo (MPRAGE) anatomical volume (TR = 2530 ms, TE = [1.64, 3.50, 5.36, 7.22 ms], TI = 1400 ms, flip angle = 7.0°, voxel resolution = 1.0 mm isotropic, FOV = 256 × 256, 176 sagittal slices) and a geometry-matched T2-weighted anatomical volume (TR = 3200 ms, TE = 454 ms, voxel resolution = 1.0 mm isotropic, FOV = 256 × 256, 176 sagittal slices) were collected prior to functional imaging.

Functional, blood oxygenation level dependent (BOLD) data were acquired using continuously-sampled, simultaneous multislice, T2*-weighted gradient-echo planar imaging (EPI) scans (TR = 750 ms, TE = 30 ms, flip angle = 90°, voxel resolution = 3.0 mm isotropic, 10% slice gap, FOV = 72 × 72, 45 slices, 5 simultaneous slices). 484 volumes were acquired during each of the two runs of language localizer, and 504 volumes during each of the two runs of the visual-spatial working

Table 1
Data analyzed (per run) for each subdivision of the localizer.

Data analyzed	Volumes	Duration	Blocks (per condition)
100 % (full dataset)	484	6:03	16 (8/8)
75 %	367	4:35	12 (6/6)
50 %	252	3:09	8 (4/4)
25 %	138	1:44	4 (2/2)

¹ <https://osf.io/qcdtb/>

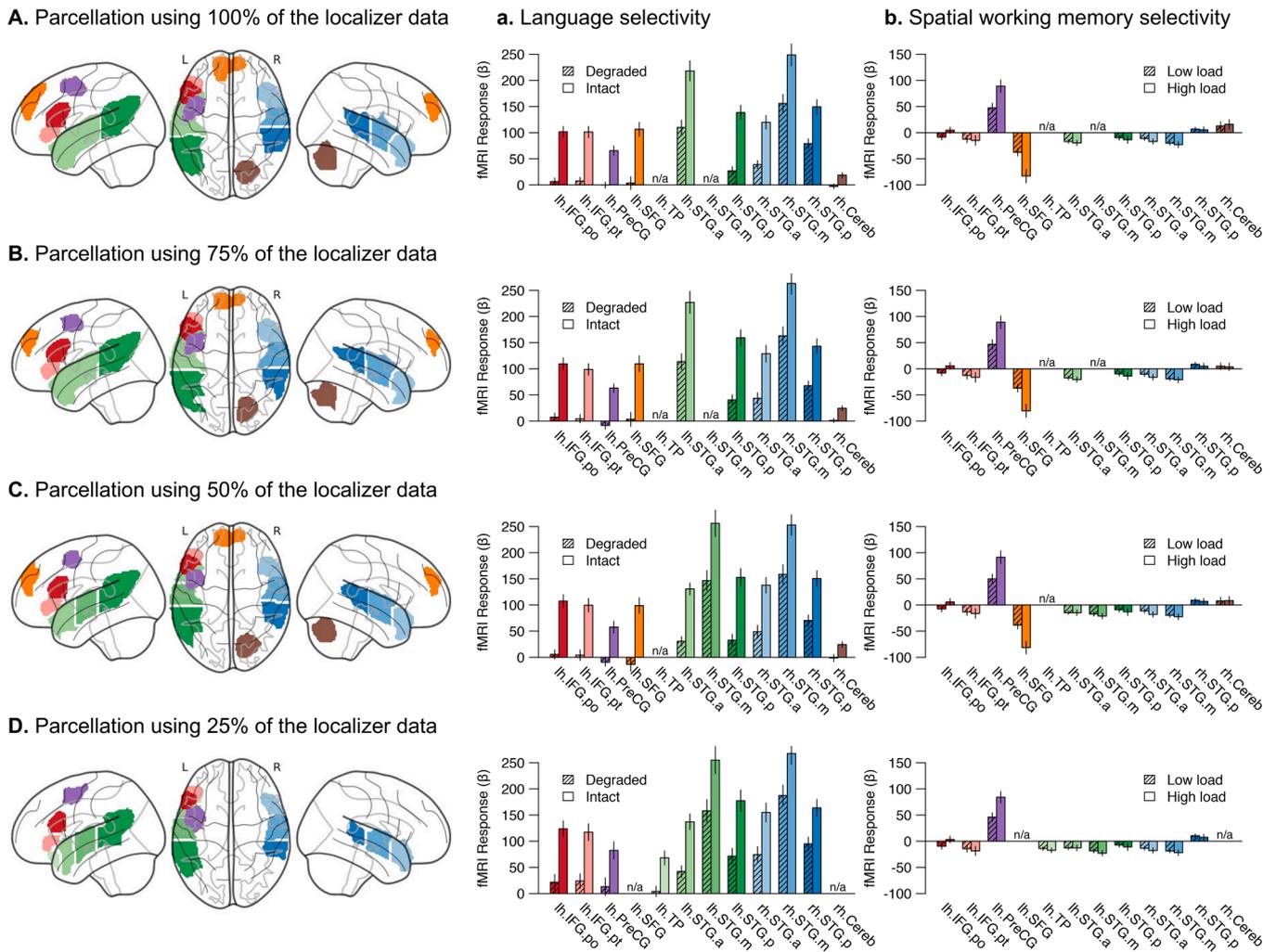


Fig. 2. Functional selectivity of language fROIs by localizer scan time. Glass brains (left column) show the location of the group-constrained parcellation based on (A) the full scan time from both runs of the localizer (16 blocks / condition and 12:06 total scan time), (B) the first 75 % of each localizer run (12 blocks / condition and 9:10 total scan time), (C) the first 50% of each localizer run (8 blocks / condition and 6:18 total scan time), and (D) the first 25 % of each localizer run (4 blocks / condition and 3:28 total scan time). From within each parcel, subject-specific fROIs were obtained and used to independently sample the fMRI response magnitude from each condition of the language localizer and spatial working memory task. The mean response magnitude across participants' fROIs within each parcel for (a) degraded vs. intact speech and (b) low vs. high working memory load is shown for each level of data subdivision. (Note that the y-axis limits differ, but the scales are the same, so magnitudes are comparable across the two tasks.) Parcellation was mostly robust to the amount of data, excepting the loss of some parcels (SFG, cerebellum) and more granular parcellation of left STG with smaller amounts of data. The pattern of selectivity across parcels was not affected by the amount of data. Parcels that were not attested for a particular localizer duration are marked "n/a."

memory task.

2.6. MRI data analysis

We subdivided each of the full 6:03- (16 block-) runs of the localizer into three additional functional series of parametrically decreasing duration (Fig. 1A; Table 1). These subdivisions comprised the first 75 % of the data from each run, the first 50 %, and the first 25 %. All subdivisions of the dataset maintained the ratio of volumes acquired during the intact and degraded speech conditions, as well as the ratio of stimulation to baseline. Data subdivisions were created from each subject's unprocessed nifti files. Functional preprocessing, within-subject modelling, and group analyses were subsequently performed for each duration of the localizer separately. (The visual-spatial working memory task data were not subdivided.)

Functional MRI data were processed using the Lyman fMRI analysis ecosystem (Waskom, 2019), in which algorithms from FSL (v5.0.7; Jenkinson et al., 2012) and FreeSurfer (v5.3.0; Dale, Fischl, and Sereno, 1999) are integrated using Nipype workflows (Gorgolewski et al., 2011).

Image preprocessing consisted of motion correction within each run (i. e., rigid-body realignment to the mean EPI image) and spatial smoothing (6mm FWHM kernel) using the SUSAN algorithm implemented in FSL (Smith and Brady, 1997). Motion and intensity outliers (functional volumes exceeding 1 mm in differential motion or differing from the mean image intensity by > 3 SD) were identified and included as nuisance regressors during modeling (Siegel et al., 2014). Model design included two task regressors for the spoken language localizer (intact and degraded speech) or two task regressors for the spatial working memory task (high-load, 6-item and low-load, 3-item sequences), as well as nuisance regressors including six motion parameters and individual regressors for any outlier volumes. Vectors for task regressors were calculated by convolving a vector of event onsets with their durations and convolving the resulting stimulation time series with a canonical hemodynamic response function to generate the hypothesized blood oxygenation level dependent response. Several contrasts of interest were computed for each participant: *intact* > *degraded* speech for the language localizer, *high* > *low* load for the spatial working memory task, as well as each of these conditions > *rest*. Within-subject estimation of the general

Table 2

Functional selectivity of language regions (intact > degraded speech contrast) was unaffected by localizer duration.

Effect	Run 1 (full) vs. Run 2 (shortened)				Run 2 (full) vs. Run 1 (shortened)			
	Estimate (β)	s.e.	t	p	Estimate (β)	s.e.	t	p
100 % vs. 75 %								
Condition (intact > degraded)	43.80	2.40	18.25	$\ll 0.001$	42.83	2.52	16.98	$\ll 0.001$
Duration (full > reduced)	6.86	2.40	2.86	< 0.005	-10.47	2.52	-4.15	$\ll 0.001$
Condition \times Duration	-1.94	2.40	-0.81	0.420	-0.44	2.52	-0.18	0.861
100 % vs. 50 %								
Condition (intact > degraded)	45.65	2.53	18.03	$\ll 0.001$	41.77	2.49	16.74	$\ll 0.001$
Duration (full > reduced)	12.17	2.59	4.70	$\ll 0.001$	-4.81	2.55	-1.89	0.059
Condition \times Duration	-3.79	2.53	-1.50	0.135	0.62	2.49	0.25	0.804
100 % vs. 25 %								
Condition (intact > degraded)	42.24	2.99	14.14	$\ll 0.001$	42.73	2.73	15.63	$\ll 0.001$
Duration (full > reduced)	1.18	3.32	0.36	0.721	-13.48	3.04	-4.43	$\ll 0.001$
Condition \times Duration	-0.38	2.99	-0.13	0.899	-0.34	2.73	-0.13	0.900

linear model and contrasts was conducted for each run in participants' native EPI space. Spatial normalization was then completed via image registration from native EPI space into common MNI-space with 2 mm isotropic voxels. First, the coregistration transformation between each participant's mean functional EPI volume and their T1-weighted structural image was calculated using Freesurfer's BBRegister program with FLIRT initialization (Greve and Fischl, 2009). Second, a nonlinear transformation was applied using ANTS v.1.9 (Avants et al., 2011) for accurate coregistration between the high-resolution structural anatomy and the MNI template. Individual subjects' statistical maps were created after transformed contrast images were combined across runs in fixed-effects analyses, as well as for each run by itself (for test-retest analyses).

2.7. Group-constrained subject-specific (GCSS) analysis

To determine the consistency of the localizer in identifying the brain regions responsive to language, a parcellation of the probabilistic location of language-responsive regions across the brain was defined at each span of data length (25 %, 50 %, 75 % and 100 % of the data from each run) using the *group-constrained subject-specific* (GCSS) approach (Fedorenko et al., 2010; Julian et al., 2012). Individual subjects' activation maps for the relevant contrast of *intact > degraded* speech were thresholded at voxelwise $p < 0.0001$. The thresholded maps were binarized and summed across subjects in common stereotaxic (MNI) space to create a probabilistic overlap map, in which each voxel encodes the number of subjects who had significant activation at that voxel for the given contrast. After smoothing the probabilistic overlap map with a Gaussian kernel of 6 mm FWHM to ameliorate variation in functional neuroanatomy across subjects and thresholding at 2 subjects (zeros assigned to voxels with fewer than 2 subjects showing suprathreshold activation), the probability map was divided into partitions, called functional *parcels*, using a watershed image segmentation algorithm that follows the topographical information in the probability map to find spatial subregions of significant activation. The key language-sensitive parcels were determined by identifying the local maxima in the probability map as the nucleus of each subregion and expanding the borders of each partition to all surrounding voxels until a local minimum between two subregions or a zero valued voxel was reached. We limited the scope of our analysis to parcels that contained significant activation from ≥ 80 % of subjects (Fedorenko et al., 2010; Julian et al., 2012).

We then obtained individual-subject *functional regions of interest* (fROIs) by intersecting each parcel from the group-level parcellation map with each individual's activation map for the relevant contrast (*intact > degraded*), sorting the voxels based on their z -values, and selecting the top 10 % of voxels within each mask as that participant's fROI. This approach ensures that each subject's fROI in a particular region contains the same number of voxels, simplifying statistical analyses performed across subjects (Nieto-Castañón & Fedorenko, 2012). For

each subdivision of the localizer data, we defined fROIs based on combined activation from both runs of the data (for comparing how the spatial location of fROIs changes as a function of localizer duration), as well as separately for each run of the language localizer task (to allow for unbiased selection of voxels for testing language-selective responses (e.g., Kriegeskorte et al., 2009), and to measure how between-run test-retest reliability of fROI location changed as a function of localizer duration).

3. Results

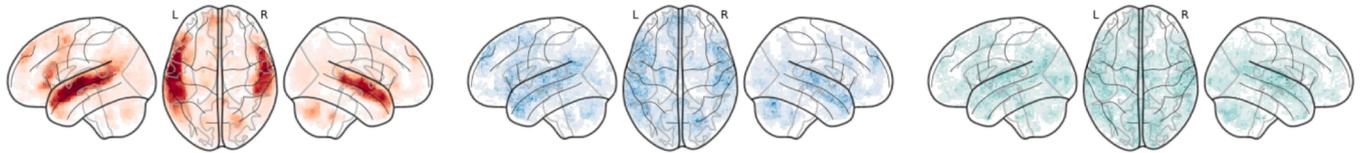
3.1. Parcellation of language regions and their functional selectivity

Application of the GCSS procedure to the auditory language localizer produced a set of language-selective functional parcels that were consistent in location, number, and extent with those reported in previous studies (Fedorenko et al., 2010; Nieto-Castañón & Fedorenko, 2012; Scott et al., 2017). The parcellation based on the full-duration localizer identified ten parcels that met the inclusion criterion: six in the left hemisphere and four in the right hemisphere (Fig. 2A). These included the anterior and posterior portions of left superior temporal gyrus (lh.STG.a, lh.STG.p), left inferior frontal gyrus pars opercularis (lh.IFG.po) and pars triangularis (lh.FG.pt), left superior frontal gyrus (lh.SFG; which also extended into the right hemisphere), left precentral gyrus (lh.PreCG), the anterior, mid and posterior portions of right STG (rh.STG.a, rh.STG.m, rh.STG.p), and right cerebellum (rh.Cereb). We repeated the GCSS procedure separately for each subdivision of the dataset. The parcellation based on 75 % of the data was essentially identical to that of the full dataset (Fig. 2B). The parcellation based on the first 50 % of the data differed only in a more granular parcellation of left anterior STG, which now consisted of separate anterior (lh.STG.a) and medial (lh.STG.m) parcels (Fig. 2C). The parcellation based on the first 25 % of the data in each run was the most different: Parcels for SFG and cerebellum did not reach the inclusion threshold, and parcellation of left anterior STG was again more granular, now including an additional parcel specific to the temporal pole (lh.TP) (Fig. 2D).

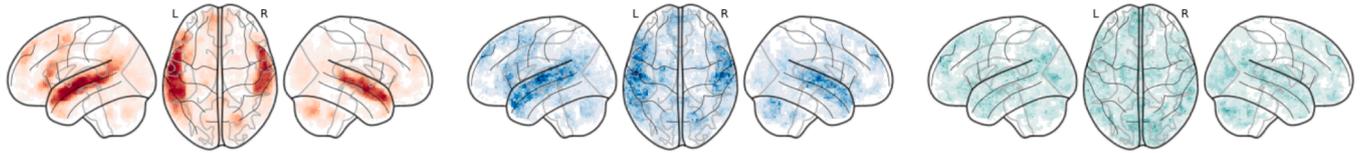
The more granular temporal lobe parcellations obtained from smaller amounts of within-subject data reflect greater heterogeneity of within-subject activation maps at the conservative voxelwise threshold (Supplementary Fig. 1A). This leads to lower likelihood of activation overlap across subjects and thus more local peaks in the group-level activation probability map from which parcels are grown (Supplementary Fig. 1B). Applying more liberal voxelwise thresholds to the smaller amounts of within-subject data yields similarly less granular parcellations, and vice-versa for more conservative thresholds of larger amounts of within-subject data (Supplementary Fig. 2).

The hallmark of a functional localizer is that it should identify neural tissue that is selectively responsive to a stimulus or cognitive operation of interest. To test whether the *selectivity* of the regions identified by the

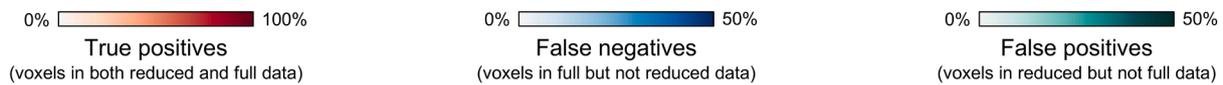
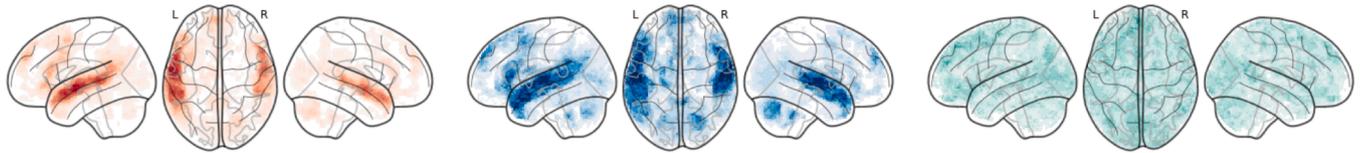
A. Whole-brain activation using 75% of the localizer data



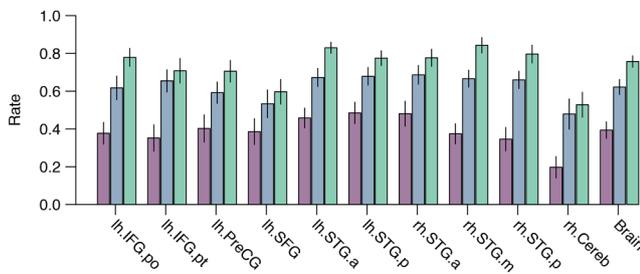
B. Whole-brain activation using 50% of the localizer data



C. Whole-brain activation using 25% of the localizer data



D. Change in true positive rate by language parcel



E. Change in false positive rate by language parcel

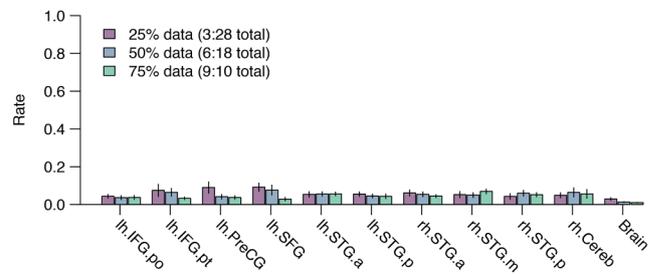


Fig. 3. Whole-brain patterns of significant voxelwise activation with reduced localizer scan time. (A-C) The proportion of subjects for whom the activation measured in each voxel during the reduced versions of the localizer reflects *true positive activation* (probability that activation for a subject in a given voxel was significant in both the full-length localizer and the corresponding reduced-length localizer; left column); *false negative activation* (probability that activation for a subject in a given voxel was significant in the full-length localizer, but not the shorter version; middle column); and *false positive activation* (probability that activation for a subject in a given voxel was significant in the reduced-length localizer, but not the full-length version; right column). Increasing scan time led to more true positives (i.e., fewer false negatives), while false positives were overall very rare. (Individual subject maps were thresholded voxelwise at $p < 0.0001$ (uncorrected); glass brains show the maximum value across the collapsed dimension for each view.) (D) Increased scan time tended to improve the true positive rate uniformly across for all probabilistic language areas. (E) The false positive rate was likewise very low across the entire brain, with only modestly more false positives found at the very shortest scan time.

language localizer differed as a function of the amount of data used to define them, we defined individual fROIs in the parcels derived from each subdivision of the data and tested the responses (fMRI activation magnitude) in these fROIs to the contrasts of interest for language and visual-spatial working memory. These values were analyzed using linear mixed-effects models, with fMRI response magnitude as the dependent variable, categorical fixed factors of *condition* (intact vs. degraded (language) or high vs. low load (spatial working memory)) and localizer *duration* (full vs. either 75 %, 50 % or 25 %), and a random effects structure including by-subject and by-parcel intercepts. To meet the models' assumption of independence, we ran a series of these comparisons, testing in each whether the fMRI response magnitudes obtained using the full-length data from one run differed from those obtained using each shorter version of the other run, and vice versa.

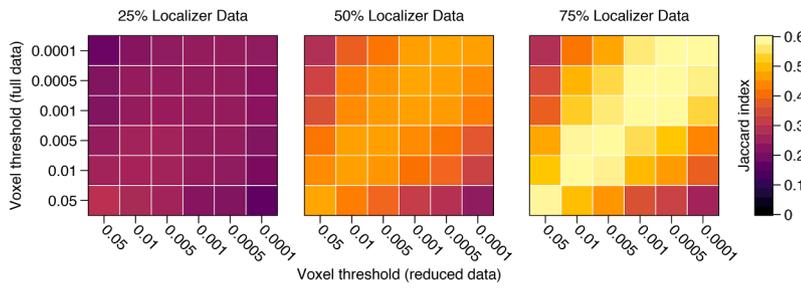
The selectivity of the fROIs obtained from each parcel for each localizer duration is shown in Fig. 2 (middle column), averaged across runs. Detailed comparisons of these data are presented in Table 2. In these models, the *condition* × *duration* interaction effect tests whether the selectivity of the language-network fROIs (effect of *condition*) was

affected by the amount of data (effect of *duration*) used to obtain the response estimates. In all comparisons against the full-length localizer, the interaction effect was not significant. That is, even when localized using just 25 % of the data (1:44 per run), the neural responses to language vs. non-language stimuli in these fROIs were just as strong as when localized using 4x as much scan time.

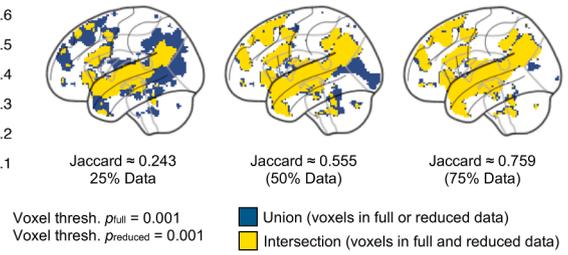
The *duration* term was also significant in many of these models, but the direction of this effect depended on which side of the contrast contained data from Run 1 vs. Run 2. Response magnitude estimates were greater for Run 1 than Run 2 (Supplementary Fig. 6); but because all subjects underwent these runs in the same order, these data cannot speak to whether this is a stimulus (e.g., speech content) or task (e.g., participant fatigue) effect. Nonetheless, comparing the two full runs against each other likewise showed no *condition* × *run* interaction, indicating both runs were equally effective at localizing language-selective voxels.

In addition to being preferentially responsive to linguistic vs. nonlinguistic stimuli, functional response in the language network is known to be insensitive to the level of difficulty of nonlinguistic,

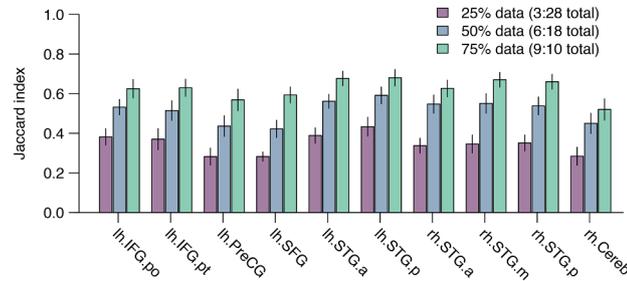
A. Whole-brain activation overlap between reduced and full data



B. Illustrative subject whole-brain activation overlap



C. fROI spatial overlap between reduced and full data



D. Illustrative subject fROI spatial overlap (lh.IFG.po)

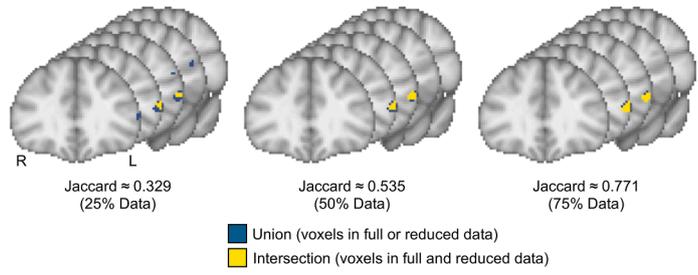


Fig. 4. Patterns of spatial convergence between the shortened localizers and the full dataset. (A) The extent of spatial overlap in whole-brain activation between the reduced localizers and the full dataset was quantified by the Jaccard index (JI) at various voxelwise thresholds (the mean value across participants is shown). (B) To help reify what the whole-brain JI values represent, here we show glass brains from a single representative subject depicting the extent of activation observed in either the reduced or the full localizer (union) compared to activation observed in both durations (intersection). (C) The spatial overlap (within participant) of the individually-defined fROIs obtained from each parcel at each reduced duration compared to those obtained from the full-length localizer. To facilitate comparison, the parcellation from the full localizer was used to define all fROIs. Bar plots show the mean across participants; error bars show \pm SEM across participants. (D) As in C, coronal slices centered on the lh.IFG.po parcel depict the location of voxels in the fROI from one representative subject in either the reduced or full localizer (union) or common to both durations (intersection). Note the impressive degree of spatial overlap captured even by JI = 0.535.

domain-general cognitive tasks (Fedorenko et al., 2011). In the corresponding models testing neural response within subject-specific language fROIs to the visual-spatial working memory task (Table 3), we found no effect of *condition*, such that the response of language fROIs overall did not differ between high vs. low spatial working memory load. Critically, there was no *condition* \times *duration* interaction for fROIs derived from any of the reduced durations vs. those based on the full dataset. That is, neural response in the language fROIs was not differentially modulated by the difficulty of the spatial working memory task, regardless of how much or how little of the localizer data had been used to define them (Fig. 2, right column). Even when localized using just 25 % of the data, these language fROIs showed no additional modulation by a multiple-demand task (Fedorenko et al., 2010; Nieto-Castañón & Fedorenko, 2012; Saxe et al., 2006).

Finally, we compared the spatial patterns and functional selectivity of the language-responsive regions identified by each of the localizer durations via classical whole-brain, group-level univariate analyses. We derived group-level fROIs from significant clusters of activation for each localizer duration formed through either cluster-level (Supplementary Fig. 3) or voxel-level (Supplementary Fig. 4) correction for multiple comparisons. In contrast to parcellations made using the GCSS analysis, patterns of activation defined by whole-brain univariate group-level contrasts were highly susceptible to effects of both localizer duration and statistical thresholding. Moreover, the functional response profiles of individuals' language regions obtained from fROIs defined by group-level activation maps were significantly less selective for the intact > degraded language contrast than those obtained using GCSS.

3.2. Spatial consistency of language regions identified by shortened localizers

As the amount of localizer data used to obtain the group-level parcellation led to only modest differences in the regions identified, and it

had no effect on the degree of task selectivity for individuals' fROIs obtained from those parcellations, we next examined the extent to which the shortened versions of the localizer identified the same neural tissue as the full-length localizer within individuals. First, we compared individual subjects' whole-brain activation maps obtained from the full localizer to those obtained from the three shorter versions. For each pairwise comparison (full vs. 75 %, full vs. 50 %, full vs. 25 %), we identified the voxels that were significantly activated in common between the shorter and full localizer (i.e., *true positives*), the voxels that were active in the full localizer but not in the shortened one (i.e., *false negatives* or Type-II errors), and the voxels that were active in the shorter localizer but not the full dataset (i.e., *false positives*, or Type-I errors) (Fig. 3). Here it is important to emphasize that the operationalization of a "true positive" voxel is made only with respect to the functional activation elicited by the full-duration localizer, and not some other ground-truth knowledge of these voxels' functional properties.

The whole-brain true-positive rate was strongly affected by the amount of localizer data (Fig. 3A–C, left column). The mean whole-brain true-positive rate across subjects improved from 39.4 % for localizer runs of 1:44, to 62.2 % for runs of 3:09, and to 75.7 % for runs of 4:35. The mean whole-brain false-positive rate across subjects was overall very low (Fig. 3A–C, right column), and, in contrast to the true-positive rate, was largely unaffected by localizer duration. The false-positive rate declined from 2.93 % (1:44 runs) to 1.27 % (3:09 runs) to 0.98 % (4:35 runs). When examining these effects separately by parcel (Fig. 3D and E) it is evident that changes in Type-I and Type-II error rates were realized similarly across the various language areas.

Because the effects shown in Fig. 3 reflect voxels localized using the same voxelwise *p*-value threshold between full and shorter localizers, we were interested in whether relaxing the statistical threshold of the shorter localizers would lead to greater convergence in the voxels identified as significant in the intact > degraded contrast. To do so, we quantified the spatial overlap between voxels localized by the full and

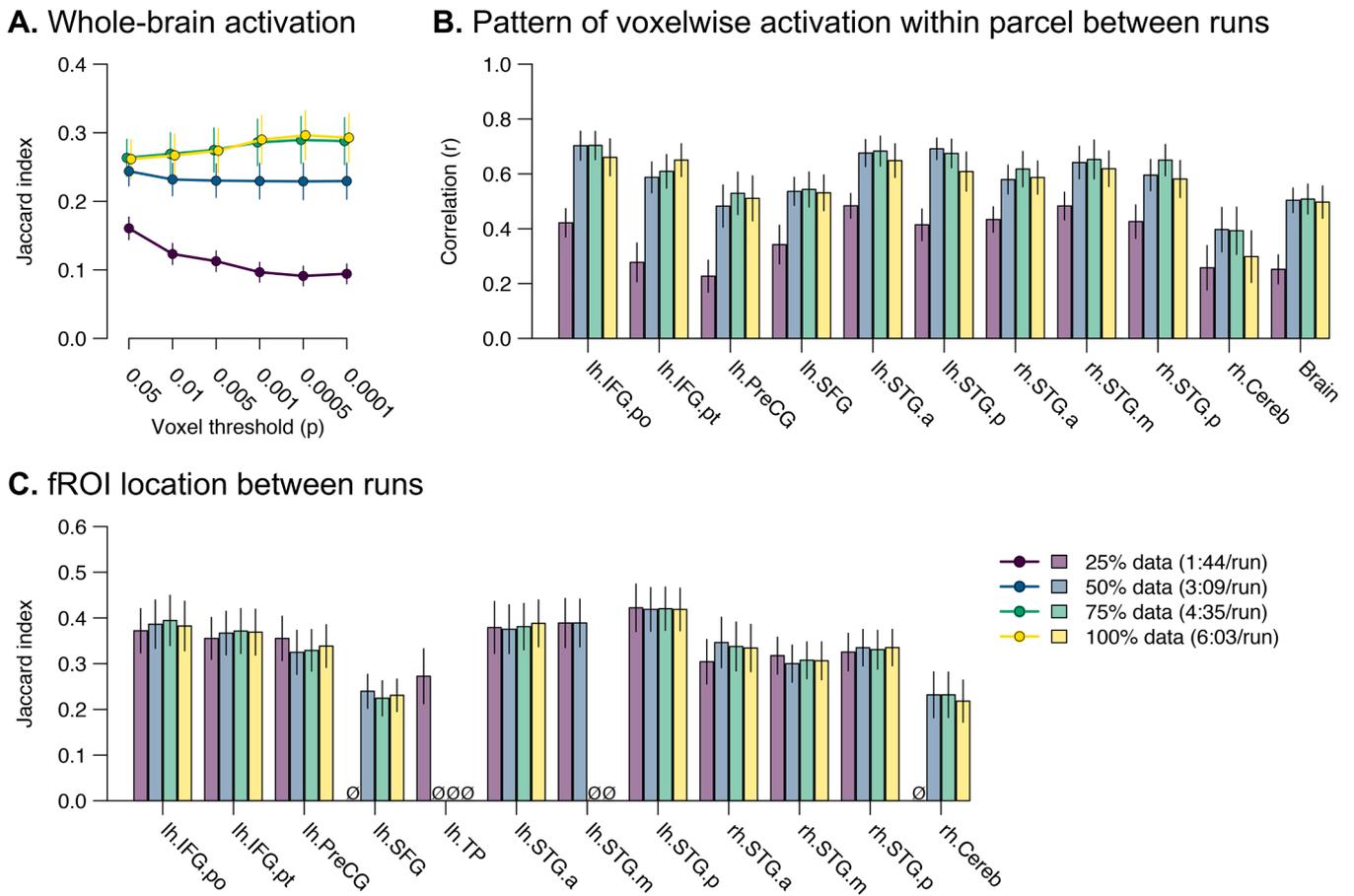


Fig. 5. Test-retest reliability between runs as a function of localizer duration. (A) The quantified spatial overlap in whole-brain activation (intact > degraded speech, at various voxelwise thresholds) between Run 1 and Run 2 for the full and reduced localizers. (B) The similarity of activation patterns within each parcel and in the whole brain between Run 1 and Run 2, operationalized as the voxelwise correlation in contrast values (intact > degraded) between runs. (C) The quantified spatial overlap (within participants) in the location of the fROIs for each parcel defined using the data from Run 1 vs. Run 2 for the full and reduced localizers. Some parcels were not found for some localizer durations (see Fig. 2), as indicated here with “∅”. For all plots, the mean across participants is shown; error bars show ± SEM across participants.

reduced localizers across the whole brain using the *Jaccard index* (JI; Jaccard, 1908) statistic, which expresses the proportion of voxels identified in common between both localizers (the contrast maps’ *intersection*) with respect to the total voxels identified by either localizer (their *union*). This metric is widely used to assess spatial overlap in functional activation maps (e.g., Maitra, 2010; Rombouts et al., 1998), as it reflects the correspondence of two samples with respect to their spatial distribution, regardless of activation magnitude (e.g., Kampa et al., 2020). To test the effect of statistical thresholding on sensitivity and specificity, we parametrically varied the voxelwise threshold of each localizer map in six steps between $p = 0.0001$ and $p = 0.05$ and calculated the whole-brain JI values for each participant at each pairwise combination of localizer thresholds. The pairwise comparisons of significant voxel overlap are illustrated in Fig. 4A. To help make the JI values more tangible, in Fig. 4B we show a single subject’s pairwise overlap maps for each localizer duration, with each input map thresholded at $p < 0.001$.

If reducing the statistical threshold of the shortened localizer led to greater overlap with the full localizer, we would have expected to see greater JI values in off-diagonal cells in the upper triangle of the matrices (i.e., where the threshold of the shortened localizer was more liberal than in the full localizer). This pattern was not observed; instead, the greatest JI values were always seen along the matrix diagonal, suggesting that statistical thresholding did not penalize the shortened localizers’ true-positive voxels. Furthermore, the spatial overlap between the voxels identified by the full localizer and either the 75

%-duration or 50 %-duration localizers tended to be greatest at the most conservative thresholds ($p = 0.001$ and below), further suggesting that more liberal statistical thresholding does not improve localizer accuracy. However, this pattern was reversed for the 25 %-duration localizer, which showed the greatest overlap with the full localizer at the most liberal p -values ($p = 0.005$ and above), which likely reflects the dramatic reduction in significant voxels found at the single-subject level for more conservative thresholds in the 25 % localizer (e.g., Supplementary Figs. 1A and 3).

In assessing the utility of shortened localizers, it is not only important to gauge their consistency in identifying language-selective activation across the whole brain, but also the extent to which they converge on identifying the peak language-selective voxels for any given individual within each parcel across localizer durations. From the parcellation based on the full localizer duration (Fig. 2A), we quantified the spatial overlap in the subject-specific fROIs obtained within each parcel for each shortened localizer duration vs. the full dataset, examining how the fROI convergence within each parcel changed as a function of localizer duration. The voxels included in fROIs defined by reduced localizers showed increasing convergence on those identified by the full localizer as the localizer durations increased (Fig. 4C). This pattern of improved spatial convergence obtained similarly across all parcels. The improvement in spatial convergence for the fROI in one parcel is shown for a representative subject (Fig. 4D).

3.3. Test-retest reliability of language regions identified by shortened localizers

In most cases, researchers will want to obtain two separate runs of a functional localizer, in order to identify fROIs and test hypotheses about their functional response profiles in independent data (Julian et al., 2012). We therefore tested the extent to which the identification of language-selective voxels (both across the whole brain, and also for the fROIs within each parcel) were reliable across two separate runs of each localizer duration. Across the whole brain, language-selective voxels were identified based on the intact > degraded contrast at voxelwise thresholds that varied in six steps from $p = 0.05$ to $p = 0.0001$, and the spatial convergence in activation between runs was quantified using the Jaccard index. Larger JI values indicate more overlap and thus more reliable whole-brain activation between runs.

The between-run test-retest reliability of whole-brain activation improved dramatically as the localizer duration increased from 1:44 to 3:09 (Fig. 5A). Whole-brain test-retest reliability improved more subtly as localizer duration increased from 3:09 to 4:35, while increasing scan time to 6:03 appeared to offer no additional improvement. For longer localizer durations, whole-brain test-retest reliability was not affected by the voxelwise p -value threshold. However, between-run activation became less reliable with more conservative thresholds for the shortest runs of 1:44.

In the GCSS technique, selecting the fROI for a parcel depends on the pattern of activation within that parcel, with the voxels that show the greatest difference in the target contrast (intact > degraded) comprising the fROI. To ascertain how the pattern of voxel activation magnitudes between runs varied as a function of localizer duration, for each subject we calculated the pairwise correlation across all voxels within each parcel between Run 1 and Run 2 of each duration of the localizer. Larger correlation coefficients indicate more similar patterns of activation between runs, and thus greater test-retest reliability of the activation within a parcel.

The between-run (test-retest) reliability of the activation pattern within each parcel improved as localizer run duration increased from 1:44 to 3:09, but the activation patterns did not become any more consistent with longer runs (Fig. 5B). This pattern appeared consistently across all parcels. However, activation in some parcels tended to have overall more reliable activation patterns (e.g., lh.IFG.po, lh.STG.a, lh.STG.p, and rh.STG.m) than did others (lh.PreCG, lh.SFG, and rh.Cereb). These variations in the pattern of activation from run to run unsurprisingly led to slight variations in group-level parcellations when obtaining these parcellations from only one run at a time (Supplementary Fig. 5), consistent with the observation of more granular parcellations with smaller amounts of within-subject data (Fig. 2).

Finally, we investigated whether the between-run likelihood of selecting particular voxels to comprise the fROI changed as a function of localizer duration. For each localizer duration, within each parcel, the spatial convergence in voxels selected for each subject's fROI was quantified using the Jaccard index. Larger JI values indicate more overlap and thus more reliable fROI sampling between runs. The between-run reliability of fROI selection was not affected by localizer duration (Fig. 5C) – a pattern observed consistently across all parcels. However, the location of fROIs had greater overall test-retest reliability in some regions (e.g., lh.STG) compared to others (e.g., lh.SFG and rh.Cereb). Notwithstanding any differences in the spatial location of the fROIs between runs, there were no between-run differences in their selectivity for either the intact vs. degraded speech contrast (Supplementary Fig. 6) or insensitivity to the hard vs. easy spatial working memory contrast (Supplementary Fig. 7).

4. Discussion

This study examined how measurement of the spatial extent and functional selectivity of language-responsive brain areas was affected by

the amount of fMRI scan time devoted to an auditory language localizer. In particular, we wanted to quantify the extent to which shorter versions of the localizer affected (i) the consistency of group-level, whole-brain parcellations of the language network; (ii) the degree of language selectivity within fROIs defined in individual subjects; (iii) the spatial consistency of language regions at the single-subject level both across the whole brain and for individually-identified fROIs; and (iv) the test-retest reliability of within-subject fMRI activation across separate runs.

While each of these metrics showed nuanced patterns of differences as a function of the amount of data, two major themes were observed: First, the brain regions identified by the localizer maintained a consistently high degree of functional selectivity for language stimuli regardless of how much (or how little) scan time was used to localize them. Second, the localizer duration could be reduced by up to 50% while having only minimal impact on essentially every metric of spatial accuracy relative to the full-duration localizer and test-retest reliability between runs.

4.1. Group-level parcellation consistency

The probabilistic parcellation of language areas at the group level was minimally affected by the duration of the localizer. The parcellations based on 75 % and the full duration localizer were effectively identical, and that based on 50 % differed only in more granular parcellation of left anterior temporal lobe. In the parcellation based on 25 % of the localizer duration, left temporal parcellation was even more granular, and two extrasylvian language parcels were not identified: SFG and right cerebellum.

The loss of SFG and cerebellar parcels with diminishing scan time was related to the threshold for inclusion in the parcellation: Compared to the a priori inclusion criterion of 80 %, significant activation was found for only 18/24 participants (75 %) in SFG and 16/24 (67 %) in right cerebellum when analyzing the shortest localizer duration. Reducing the parcel inclusion criterion to 67 % would have preserved these two parcels without adding any new ones not identified by the longer durations. It is worth noting that the criterion for parcel inclusion is not absolute, being one of the many discretionary thresholds that researchers must set when carrying out fMRI, including GCSS, analyses (Kawabata Duncan & Devlin, 2011).

Meanwhile, more granular parcellation of anterior left temporal lobe with shorter localizers may actually reflect false negatives identifying parcel boundaries based on larger amounts of data. Original studies using GCSS-based parcellation of the language network also reported four parcels in left temporal lobe (Fedorenko et al., 2010). It is likely that the present paradigm was more effective at measuring language activation in this part of the brain, owing to the use of auditory stimulation and a shorter TR (i.e., more within-subject samples; Nee, 2019) than these earlier reports, leading to better within-subject detection of language-selective cortex in anterior temporal lobe. Setting even more conservative thresholds for within-subject activation may lead to similarly granular temporal lobe parcellations when using longer scan durations, and more liberal thresholds for shorter scan durations may both reduce granularity and increase the inclusion of other marginally representative parcels.

Ultimately, questions about the quality of parcellations obtained using only a small amount of within-subject data may be moot given the availability of open-source parcellations based on large numbers of subjects (e.g., Lipkin et al., 2022). If a study's goal is to test the responsiveness of individual-subject fROIs within the language network, such an analysis does not require defining parcels using one's own data. Indeed, the definition of a parcel in a GCSS analysis is a region where activation is likely to be found within a population. In this way, using an a priori parcellation can allow the unbiased selection of language fROIs in small- N and case-study fMRI experiments that would otherwise be underpowered to obtain group-level parcellations from their own data.

4.2. Language selectivity of individual fROIs

Perhaps the most important observation from these data is that specialization for linguistic processing remained consistent across the language network regardless of the amount of scan time. Using an unbiased approach to independently select and test the response profiles of subject-specific fROIs in the language network, we found a strikingly consistent pattern of selectivity for intact vs. degraded speech and insensitivity to high vs. low spatial working memory load. Even when language and working memory response selectivity was measured based on only 1:44 of scan time, all nodes of the language network showed equivalent response profiles as when measured using more than three times as much data.

These results demonstrate that robustly language-selective fROIs can be identified in the brains of individuals with less than two minutes of scan time. Because identification of these fROIs can be achieved so expediently, this leaves substantially more scan time available to researchers for collecting data on other, hypothesis-driven questions about the response properties of the language network, while gaining access to all of the theoretical and statistical advantages that subject-specific fROIs offer over traditional group analyses (Saxe et al., 2006; Fedorenko, 2021).

4.3. Spatial consistency of language regions in individual subjects

While the conservatively defined fROIs remained equally language-selective, analyzing smaller amounts of data also always led to fewer true positive voxels (vs. the full duration localizer) across the entire brain. This reduction was not linear, with greater reductions in true positive voxel identification when shortening the localizer from 100 % to 75 % and from 50 % to 25 %, with a much more modest effect between the 75 and 50 % durations. False negatives were equally likely across frontal, temporal, and extrastriatal areas, such that no language regions were more or less affected when scan time was reduced. Importantly, the whole-brain false positive rate was always very low and was unaffected by localizer duration, indicating that even the shortest localizer is unlikely to risk introduction of theoretically spurious measurements.

One might have imagined that setting liberal within-subject voxel thresholds for shorter-duration localizers would have yielded whole-brain maps that were more similar to conservative thresholds at longer durations. However, this turned out not to be the case. The within-subject consistency in whole-brain language maps tended to be greatest between localizer durations when the voxelwise threshold was (i) the same for both durations, and (ii) more conservative (i.e., $p < 0.001$ and smaller). This counterintuitive result suggests that maximally language-responsive areas are identified with high reliability regardless of paradigm duration, but that there is more variable detection of peripheral activations in all cases. This finding parallels the observation above that the fROIs (i.e., the most selective voxels) are always highly language selective regardless of localizer duration.

4.4. Within-subject test-retest reliability

While the primary purpose of a localizer is to define fROIs to be applied to additional, hypothesis-driven contrasts, validating the effectiveness of a localizer nonetheless requires independent training (fROI-defining) and testing (response-measuring) data. Often this is achieved through measuring two runs using different stimuli. Measuring the within-subject test-retest reliability of the localizer across runs revealed a number of interesting observations: First, the degree of between-run overlap in language-selective voxels across the whole brain was essentially the same regardless of localizer duration and voxelwise threshold in all cases except when analyzing the shortest (25 %) duration localizer. The between-run activation pattern (i.e., relative response magnitude per voxel) within language parcels was also essentially the same for all

localizer durations except the shortest. Furthermore, the degree of spatial overlap for fROIs defined on each run of data was unaffected by localizer duration. These results suggest that, for a given amount of scan time, the map of language selective voxels produced by the intact > degraded speech contrast is highly reliable within an individual subject. While there was more overall test-retest variability in this map across the whole brain at the shortest scan time, the areas of peak activation (i.e., those selected for fROIs) were nonetheless equally reliable vs. longer scanning durations. These results further reveal the robustness of even very short versions of the localizer for selecting valid and reliable language-selective fROIs within individual subjects.

4.5. Recommendations for practice

Heuristically, more data are always better. But the acquisition of neuroimaging data is costly in terms of both monetary expense and human time and effort. Pragmatic and economic considerations necessarily impose trade-offs on how much data can be collected during any particular fMRI session or task. For instance, while longer localizers may lead to more accurate localization in ideal situations, they also increase the likelihood of participant noncompliance when working with children and other special populations, which could vitiate the localizer entirely. How, then, can researchers use these results to guide them in adding language localizers to their own fMRI studies?

Perhaps the most important consideration in selecting the amount of localizer scan time is the purpose for localizing language-selective regions. One common use for functional language localizers has been to examine whether activity in the core language network is dissociable from other mental operations (e.g., Blank et al., 2014; Mineroff et al., 2018) and whether various linguistic operations can be differentially associated individual fROIs (Fedorenko et al., 2015; Siegelman et al., 2019). If the goal is to characterize the functional signature of nodes of the brain's language network, then the present data suggest that localizer scan time can in fact be drastically reduced. A single run of 1:44 yields language-selective fROIs that appear to be just as selective for language and insensitive to domain-general cognitive tasks as those obtained after two runs totaling 12:06. These fROIs can be selected from parcels defined on large, independent samples (Lipkin et al., 2022), further reducing concerns about not collecting either enough within- or between-subjects data in any one study to define a set of group-constrained parcels. If researchers do want to construct a parcellation based on their own sample – for instance, if testing hypotheses about differences in language network organization between groups or ages (Hiersche et al., 2022; Lee, 2022) – then the shortest localizer may not be appropriate. From the present data, whole-brain language activation maps were most reliable with localizer durations of 50% (two runs of 3:09) or more.

However, if the localizer is intended to comprehensively demarcate language-selective cortex (i.e., to minimize the spatial extent of false negatives), then it is instead advisable to obtain as much data as possible. For instance, there is hope that fMRI can one day serve the needs of presurgical mapping (e.g., Diachek et al., 2022; Wilson et al., 2017), where minimization of false negatives is paramount. In such cases, shorter localizers may be convenient but risk greater human cost. Current research on the language network has not identified the necessary vs. sufficient relationship between the degree of language response and functional language outcomes after injury or surgery. The present study did not attempt to identify an upper bound on the amount of scan time after which no further gains in language mapping were observed, but that number likely exceeds the longest measurements we employed.

Finally, we should recall that the present observations are based on comparisons of data obtained from a single localizer paradigm. This raises two important caveats regarding interpretation and generalization of these conclusions: First, inferential statistical comparisons assume independence of measurement errors, which is violated when

Table 3

Language-selective fROIs were insensitive to domain-general task demands (high > low spatial working memory load) regardless of the localizer duration used to define them.

Effect	Run 1 (full) vs. Run 2 (shortened)				Run 2 (full) vs. Run 1 (shortened)			
	Estimate (β)	s.e.	t	p	Estimate (β)	s.e.	t	p
100 % vs. 75 %								
Condition (high load > low load)	-0.47	1.45	-0.32	0.748	-0.39	1.43	-0.28	0.783
Duration (full > reduced)	-0.32	1.45	-0.22	0.824	0.93	1.43	0.65	0.514
Condition \times Duration	0.25	1.45	0.17	0.863	-0.06	1.43	-0.04	0.968
100 % vs. 50 %								
Condition (high load > low load)	-0.42	1.40	-0.30	0.766	-0.37	1.39	-0.27	0.790
Duration (full > reduced)	-0.83	1.43	-0.58	0.564	0.65	1.42	0.46	0.646
Condition \times Duration	0.20	1.40	0.14	0.885	-0.08	1.39	-0.06	0.954
100 % vs. 25 %								
Condition (high load > low load)	0.37	1.35	0.27	0.784	0.54	1.30	0.42	0.677
Duration (full > reduced)	-0.85	1.50	-0.57	0.569	0.43	1.45	0.30	0.765
Condition \times Duration	-0.59	1.35	-0.43	0.665	-0.99	1.30	-0.76	0.447

comparing partially overlapping samples. This can underestimate true error and overestimate significance of nonindependent effects (e.g., differences between measurements obtained from the reduced vs. full localizer). In some cases, we have been able to maintain independence by comparing strictly nonoverlapping samples from two separate runs (as in Section 3.1). In other cases, we have eschewed inferential statistical tests when the relevant comparisons can be made qualitatively and the descriptive statistics speak for themselves (as in Sections 3.2 and 3.3). We opted for this design because it better reflects how researchers actually evaluate design considerations when planning new neuroimaging experiments: Assuming participants will all undergo the same imaging paradigm, how much data does an experiment need to collect? Second, and relatedly, we cannot claim that every stimulation paradigm can effectively localize the language network in under two minutes, only that it is possible both in principle and when using this paradigm (Scott et al., 2017). Researchers planning new paradigms will want to ensure the effectiveness of their stimuli and task, and we hope this manuscript provides a framework for doing so.

5. Conclusions

Using a localizer to demarcate language-selective fROIs in individual brains offers considerable theoretical advantages for investigating the functional organization of the language network. Here, we found that core nodes of the language network exhibited highly language-selective responses even when localized using as little as 1:44 of scan time. The spatial extent of these regions at the group level, and the locations of subject-specific activation peaks, could be reliably identified using localizer runs of three to four minutes each. These results demonstrate the high degree of focal selectivity of the cortical language network. Correspondingly, they demonstrate that, for many cognitive neuroscience applications, a functional localization of individual subjects' language-selective areas can be effectively implemented with much less scan time and much less cost than in most prior studies. These findings create new opportunities to use a functional language localizer to study the neural organization of language in special populations.

Data availability statement

All of the stimuli, stimulus delivery code, participant data, and analysis code are available via our institutional archive, which resides permanently at this address: <https://open.bu.edu/handle/2144/16460>.

CRediT authorship contribution statement

Jayden J. Lee: Methodology, Software, Formal analysis, Writing – original draft, Visualization. **Terri L. Scott:** Conceptualization, Methodology, Software, Investigation, Data curation, Writing – review &

editing, Visualization. **Tyler K. Perrachione:** Conceptualization, Software, Validation, Formal analysis, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

None.

Data availability

Stimuli, stimulus delivery code, participant data, and analysis code are available via our institutional archive, which resides permanently at this address: <https://open.bu.edu/handle/2144/16460>.

Acknowledgments

We thank Jessica Tin, Yaminah Carter, and Ja Young Choi for their assistance; Ev Fedorenko, Frank Guenther, and Emily Stephen for helpful discussion; and Atsushi Takahashi, Steve Shannon, and Sheeba Arnold at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research, MIT for technical support during scanning. Research reported in this article was supported by the National Institute on Deafness and Other Communications Disorders (NIDCD) at the National Institutes of Health (NIH) under award number R03 DC014045 and a NARSAD Young Investigator Grant to TP. JL was supported by NIDCD training grant T32 DC013017. TS was supported by the National Institute on Drug Abuse (NIDA) training grant T90 DA032484 and NIDCD under award number F32 DC019531.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.neuroimage.2023.120489](https://doi.org/10.1016/j.neuroimage.2023.120489).

References

- Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* 54 (3), 2033–2044. <https://doi.org/10.1016/j.neuroimage.2010.09.025>.
- Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. *Nature* 403 (6767), 309–312. <https://doi.org/10.1038/35002078>.
- Berman, M.G., Park, J., Gonzalez, R., Polk, T.A., Gehrke, A., Knaffla, S., Jonides, J., 2010. Evaluating functional localizers: the case of the FFA. *Neuroimage* 50 (1), 56–71. <https://doi.org/10.1016/j.neuroimage.2009.12.024>.
- Blank, I., Kanwisher, N., Fedorenko, E., 2014. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *J. Neurophysiol.* 112 (5), 1105–1118. <https://doi.org/10.1152/jn.00884.2013>.

- Brett, M., Johnsrude, I.S., Owen, A.M., 2002. The problem of functional localization in the human brain. *Nat. Rev. Neurosci.* 3 (3), 243–249. <https://doi.org/10.1038/nrn756>.
- Broca, P., 1861. Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (parte de la parole). *Bull. Soc. Anat. Paris* 6, 330–357.
- Cohen, L., Lehericy, S., Chochon, F., Lemer, C., Rivaud, S., Dehaene, S., 2002. Language-specific tuning of visual cortex? Functional properties of the visual word form area. *Brain* 125 (Pt 5), 1054–1069. <https://doi.org/10.1093/brain/awf094>.
- Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* 9 (2), 179–194.
- Dehaene, S., Le Clec, H.G., Poline, J.B., Le Bihan, D., Cohen, L., 2002. The visual word form area: a prelexical representation of visual words in the fusiform gyrus. *Neuroreport* 13 (3), 321–325. <https://doi.org/10.1097/00001756-200203040-00015>.
- Diachek, E., Blank, I., Siegelman, M., Affourtit, J., Fedorenko, E., 2020. The domain-general multiple demand (MD) network does not support core aspects of language comprehension: a large-scale fMRI investigation. *J. Neurosci.* 40 (23), 4536–4550. <https://doi.org/10.1523/JNEUROSCI.2036-19.2020>.
- Diachek, E., Morgan, V.L., Wilson, S.M., 2022. Adaptive language mapping paradigms for presurgical language mapping. *AJNR Am. J. Neuroradiol.* 43 (10), 1453–1459. <https://doi.org/10.3174/ajnr.A7629>.
- 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. <https://doi.org/10.1126/science.1063414>.
- Elliott, M.L., Knodt, A.R., Hariri, A.R., 2021. Striving toward translation: strategies for reliable fMRI measurement. *Trends Cogn. Sci.* 25 (9), 776–787. <https://doi.org/10.1016/j.tics.2021.05.008>.
- Fassbender, C., Mukherjee, P., Schweitzer, J.B., 2017. Minimizing noise in pediatric task-based functional MRI: Adolescents with developmental disabilities and typical development. *Neuroimage* 149, 338–347. <https://doi.org/10.1016/j.neuroimage.2017.01.021>.
- Fedorenko, E., 2021. The early origins and the growing popularity of the individual-subject analytic approach in human neuroscience. *Current Opinion in Behavioral Sciences* 40, 105–112.
- Fedorenko, E., Behr, M.K., Kanwisher, N., 2011. Functional specificity for high-level linguistic processing in the human brain. *Proc. Natl. Acad. Sci. U. S. A.* 108 (39), 16428–16433. <https://doi.org/10.1073/pnas.1112937108>.
- Fedorenko, E., Blank, I.A., Siegelman, M., Mineroff, Z., 2020. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition* 203, 104348. <https://doi.org/10.1016/j.cognition.2020.104348>.
- Fedorenko, E., Duncan, J., Kanwisher, N., 2012. Language-selective and domain-general regions lie side by side within Broca's area. *Curr. Biol.* 22 (21), 2059–2062. <https://doi.org/10.1016/j.cub.2012.09.011>.
- Fedorenko, E., Hsieh, P.J., Balewski, Z., 2015. A possible functional localizer for identifying brain regions sensitive to sentence-level prosody. *Lang. Cogn. Neurosci.* 30 (1–2), 120–148. <https://doi.org/10.1080/01690965.2013.861917>.
- Fedorenko, E., Hsieh, P.J., Nieto-Castanon, 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. <https://doi.org/10.1152/jn.00032.2010>.
- Fedorenko, E., Varley, R., 2016. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Ann. N. Y. Acad. Sci.* 1369 (1), 132–153. <https://doi.org/10.1111/nyas.13046>.
- Fox, C.J., Iaria, G., Barton, J.J., 2009. Defining the face processing network: optimization of the functional localizer in fMRI. *Hum. Brain Mapp.* 30 (5), 1637–1651. <https://doi.org/10.1002/hbm.20630>.
- Friston, K.J., Rotshtein, P., Geng, J.J., Sterzer, P., Henson, R.N., 2006. A critique of functional localisers. *Neuroimage* 30 (4), 1077–1087. <https://doi.org/10.1016/j.neuroimage.2005.08.012>.
- Gabrieli, J.D., 2009. Dyslexia: a new synergy between education and cognitive neuroscience. *Science* 325 (5938), 280–283. <https://doi.org/10.1126/science.1171999>.
- Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., Ghosh, S.S., 2011. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinform.* 5, 13. <https://doi.org/10.3389/fninf.2011.00013>.
- Gorgolewski, K.J., Storkey, A.J., Bastin, M.E., Whittle, I., Pernet, C., 2013. Single subject fMRI test-retest reliability metrics and confounding factors. *Neuroimage* 69, 231–243. <https://doi.org/10.1016/j.neuroimage.2012.10.085>.
- Greene, D.J., Black, K.J., Schlaggar, B.L., 2016. Considerations for MRI study design and implementation in pediatric and clinical populations. *Dev. Cogn. Neurosci.* 18, 101–112. <https://doi.org/10.1016/j.dcn.2015.12.005>.
- Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* 48 (1), 63–72.
- Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8 (5), 393–402. <https://doi.org/10.1038/nrn2113>.
- Hiersche, K.J., Schettini, E., Li, J., Saygin, Z.M., 2022. Functional dissociation of the language network and other cognition in early childhood. *bioRxiv* 503597. <https://doi.org/10.1101/2022.08.11.503597>.
- Jaccard, P., 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaudoise Sci. Nat.* 44, 223–270.
- Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. FSL. *Neuroimage* 62 (2), 782–790. <https://doi.org/10.1016/j.neuroimage.2011.09.015>.
- 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. <https://doi.org/10.1016/j.neuroimage.2012.02.055>.
- Kampa, M., Schick, A., Sebastian, A., Wessa, M., Tuscher, O., Kalisch, R., Yuen, K., 2020. Replication of fMRI group activations in the neuroimaging battery for the mainz resilience project (MARP). *Neuroimage* 204, 116223. <https://doi.org/10.1016/j.neuroimage.2019.116223>.
- 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. <https://www.ncbi.nlm.nih.gov/pubmed/9151747>.
- Kawabata Duncan, K.J., Devlin, J.T., 2011. Improving the reliability of functional localizers. *Neuroimage* 57 (3), 1022–1030. <https://doi.org/10.1016/j.neuroimage.2011.05.009>.
- 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. <https://doi.org/10.1038/nn.2303>.
- Lee, J.J., 2022. Efficient localization of the cortical language network and its functional neuroanatomy dyslexia. In: [Doctoral dissertation, Boston University]. Boston University Institutional Repository. <https://hdl.handle.net/2144/43724>.
- Lipkin, B., Tuckute, G., Affourtit, J., et al., 2022. Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Sci. Data* 9, 529. <https://doi.org/10.1038/s41597-022-01645-3>.
- Maitra, R., 2010. A re-defined and generalized percent-overlap-of-activation measure for studies of fMRI reproducibility and its use in identifying outlier activation maps. *Neuroimage* 50 (1), 124–135. <https://doi.org/10.1016/j.neuroimage.2009.11.070>.
- Manca, A., Martinez, G., Cugusi, L., Dragone, D., Dvir, Z., Deriu, F., 2017. The surge of predatory open-access in neurosciences and neurology. *Neuroscience* 353, 166–173. <https://doi.org/10.1016/j.neuroscience.2017.04.014>.
- Meissner, T.W., Walbrin, J., Nordt, M., Koldewyn, K., Weigelt, S., 2020. Head motion during fMRI tasks is reduced in children and adults if participants take breaks. *Dev. Cogn. Neurosci.* 44, 100803. <https://doi.org/10.1016/j.dcn.2020.100803>.
- Mineroff, Z., Blank, I.A., Mahowald, K., Fedorenko, E., 2018. A robust dissociation among the language, multiple demand, and default mode networks: Evidence from inter-region correlations in effect size. *Neuropsychologia* 119, 501–511. <https://doi.org/10.1016/j.neuropsychologia.2018.09.011>.
- Mumford, J.A., 2012. A power calculation guide for fMRI studies. *Soc. Cogn. Affect. Neurosci.* 7 (6), 738–742. <https://doi.org/10.1093/scan/nss059>.
- Nee, D.E., 2019. fMRI replicability depends upon sufficient individual-level data. *Commun. Biol.* 2, 130. <https://doi.org/10.1038/s42003-019-0378-6>.
- Nieto-Castanon, A., Fedorenko, E., 2012. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *Neuroimage* 63 (3), 1646–1669. <https://doi.org/10.1016/j.neuroimage.2012.06.065>.
- Pernet, C.R., McAleer, P., Latinus, M., Gorgolewski, K.J., Charest, I., Bestelmeyer, P.E., Watson, R.H., Fleming, D., Crabbe, F., Valdes-Sosa, M., Belin, P., 2015. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage* 119, 164–174. <https://doi.org/10.1016/j.neuroimage.2015.06.050>.
- Perrachione, T.K., Ghosh, S.S., 2013. Optimized design and analysis of sparse-sampling fMRI experiments. *Front. Neurosci.* 7, 55. <https://doi.org/10.3389/fnins.2013.00055>.
- 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 (2), 816–847. <https://doi.org/10.1016/j.neuroimage.2012.04.062>.
- Rombouts, S.A., Barkhof, F., Hoogenraad, F.G., Sprenger, M., Scheltens, P., 1998. Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. *Magn. Reson. Imaging* 16 (2), 105–113. [https://doi.org/10.1016/s0730-725x\(97\)00253-1](https://doi.org/10.1016/s0730-725x(97)00253-1).
- Ross, P., de Gelder, B., Crabbe, F., Grosbras, M.H., 2020. A dynamic body-selective area localizer for use in fMRI. *MethodsX* 7, 100801. <https://doi.org/10.1016/j.mex.2020.100801>.
- Saxe, R., Brett, M., Kanwisher, N., 2006. Divide and conquer: a defense of functional localizers. *Neuroimage* 30 (4), 1088–1096. <https://doi.org/10.1016/j.neuroimage.2005.12.062>.
- Scott, T.L., 2020. Neural bases of phonological working memory. [Doctoral dissertation Boston University]. Boston University Institutional Repository. <https://hdl.handle.net/2144/41116>.
- Scott, T.L., Galle, J., Fedorenko, E., 2017. A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cogn. Neurosci.* 8 (3), 167–176. <https://doi.org/10.1080/17588928.2016.1201466>.
- Scott, T.L., Perrachione, T.K., 2019. Common cortical architectures for phonological working memory identified in individual brains. *NeuroImage* 202, 116096. <https://doi.org/10.1016/j.neuroimage.2019.116096>.
- Siegel, J.S., Power, J.D., Dubis, J.W., Vogel, A.C., Church, J.A., Schlaggar, B.L., Petersen, S.E., 2014. Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. *Hum. Brain Mapp.* 35 (5), 1981–1996.
- Siegelman, M., Blank, I.A., Mineroff, Z., Fedorenko, E., 2019. An attempt to conceptually replicate the dissociation between syntax and semantics during sentence comprehension. *Neuroscience* 413, 219–229. <https://doi.org/10.1016/j.neuroscience.2019.06.003>. Aug 10.
- Smith, S.M., Brady, J.M., 1997. SUSAN: A new approach to low level image processing. *Int. J. Comput. Vision* 23 (1), 45–78.
- Somers, D.C., Michalka, S.W., Tobbyne, S.M., Noyce, A.L., 2021. Individual subject approaches to mapping sensory-biased and multiple-demand regions in human frontal cortex. *Curr. Opin. Behav. Sci.* 40, 169–177. <https://doi.org/10.1016/j.cobeha.2021.05.002>.

- Stoppelman, N., Harpaz, T., Ben-Shachar, M., 2013. Do not throw out the baby with the bath water: choosing an effective baseline for a functional localizer of speech processing. *Brain Behav.* 3 (3), 211–222. <https://doi.org/10.1002/brb3.129>.
- M.L. Waskom (2019). Lyman: A Python fMRI analysis ecosystem. Available online: <https://github.com/mwaskom/lyman/tree/1.0>.
- Wilson, S.M., Bautista, A., Yen, M., Lauderdale, S., Eriksson, D.K., 2017. Validity and reliability of four language mapping paradigms. *Neuroimage Clin.* 16, 399–408. <https://doi.org/10.1016/j.nicl.2016.03.015>.