Journal of Neurolinguistics 68 (2023) 101163

Contents lists available at ScienceDirect

Journal of Neurolinguistics

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

A systematic review of neuroimaging approaches to mapping language in individuals

Aahana Bajracharya [a], Jonathan E. Peelle [b,*]

[a] Department of Radiology, Washington University in Saint Louis, United States
[b] Center for Cognitive and Brain Health, Department of Communication Sciences and Disorders, and Department of Psychology, Northeastern University, United States

ARTICLE INFO

Keywords:
Systematic review
Language
Speech
Reproducibility
fMRI
Neuroimaging
Language mapping

ABSTRACT

Although researchers often rely on group-level fMRI results to draw conclusions about the neurobiology of language, doing so without accounting for the complexities of individual brains may reduce the validity of our findings. Furthermore, understanding brain organization in individuals is critically important for both basic science and clinical translation. To assess the state of single-subject language localization in the functional neuroimaging literature, we carried out a systematic review of studies published through April 2020. Out of 977 papers identified through our search, 121 met our inclusion criteria for reporting single-subject fMRI results (fMRI studies of language in adults that report task-based single-subject statistics). Of these, 20 papers reported using a single-subject test-retest analysis to assess reliability. Thus, we found that a relatively modest number of papers reporting single-subject results quantified single-subject reliability. These varied substantially in acquisition parameters, task design, and reliability measures, creating significant challenges for making comparisons across studies. Future endeavors to optimize the localization of language networks in individuals will benefit from the standardization and broader reporting of reliability metrics for different tasks and acquisition parameters.

1. Introduction

Historically, much of our understanding of the neurobiology of language has come from lesion studies and the differing profiles of patients with acquired aphasia. This patient-centered approach necessitated some degree of relating behavior to brain damage in individual people. Subsequent advances in functional neuroimaging methods have, helpfully, broadened our view of the brain regions involved in language processing by allowing language function to be investigated in healthy brains in the absence of impairment. The majority of functional neuroimaging studies are focused on group results: that is, population inference based on a sample of individuals drawn from that population. Thus, the contemporary cognitive neuroscience of language has benefitted from the wide availability of functional neuroimaging methods and statistical approaches for group statistics.

At the same time, there continues to be an awareness that the organization of language regions in the brain differs from person to person. Cognizance of the importance of individual brain organization is perhaps nowhere more evident than in the context of presurgical planning for brain surgery and the importance of avoiding damage to regions supporting speech and language processing. The gold standard for presurgical mapping continues to be electrical stimulation mapping, in which neurosurgeons stimulate areas of

* Corresponding author.
E-mail address: j.peelle@northeastern.edu (J.E. Peelle).

https://doi.org/10.1016/j.jneuroling.2023.101163
Received 21 April 2023; Received in revised form 5 August 2023; Accepted 11 August 2023
Available online 16 August 2023

cortex and assess any gross impairments to language function. Variability in the regions of the brain supporting language function necessitates this mapping be done in each patient.

Beyond presurgical mapping, numerous clinical studies have also shown evidence for atypical language organization due to various conditions, including epilepsy (Baciu et al., 2003; Gould et al., 2016; Lee et al., 2008), aphasia (Khateb et al., 2004), vascular malformations (Hakyemez et al., 2006; Pouratian, Bookheimer, Rex, Martin, & Toga, 2002), brain injury and long-standing tumors (Avramescu-Murphy et al., 2017; Kośla et al., 2015; Partovi et al., 2012; Ruff et al., 2008), and sensory deficits caused by congenital blindness (Röder, Stock, Bien, Neville, & Rösler, 2002; Roland et al., 2013). Non-clinical characteristics such as left-handedness have been associated with an increased incidence of atypical language dominance (Acioly et al., 2014). Additionally, language processing differs for primary and secondary languages (Dehaene et al., 1997; Polczynska et al., 2016, 2017; Tomasino et al., 2014), and multilingual individuals also often demonstrate the recruitment of additional brain areas for language switching (Sierpowska et al., 2013; Tomasino et al., 2014). Although many of these studies focus on group-level results, the findings have implications for understanding brain organization in single subjects.

All these trends come together in the current resurgence of interest in mapping function in the brains of individual people and identifying reliable differences in functional connectivity in highly sampled individuals (Gordon et al., 2017; Gratton et al., 2020). In considering language organization in particular, Fedorenko and colleagues have used a group-constrained functional localizer approach (Fedorenko, Hsieh, Nieto-Castañón, Whitfield-Gabrieli, & Kanwisher, 2010) to investigate a number of aspects of language using subject-specific regions of interest (ROIs) (Diachek, Blank, Siegelman, Affourtit, & Fedorenko, 2020; Fedorenko et al., 2011, 2012). In particular, they have demonstrated that group statistics relying on atlas-based normalization can provide different results from subject-specific functional localizers. Identifying subject-specific patterns of organization appears to be one of our best tools for advancing both scientific understanding and clinical translation.

A related and long-standing concern in cognitive neuroscience has been the degree to which fMRI-based activations, generally, are valid and reliable (Bennett & Miller, 2010; Elliott et al., 2020; Gorgolewski, Storkey, Bastin, Whittle, & Pernet, 2013; McGonigle, 2012; McGonigle et al., 2000; Noble et al., 2019, 2022; Smith et al., 2005). There are numerous challenges to reliability in fMRI (for a review, see Bennett & Miller, 2010). In brief, these include MRI acquisition parameters (field strength, voxel size, etc.), analysis pipelines (Esteban et al., 2019; Zhang et al., 2009), participant movement and how to correct for it (Jones, Zhu, Bajracharya, Luor, & Peelle, 2022), and state-based physiological changes (for example, time-of-day effects). Because we do not have ground truth knowledge about language organization, validity is often established through comparison with other methods. For example, language lateralization via fMRI can be compared to that achieved with a Wada test, where agreement is typically high for people with left-lateralized language (Bauer, Reitsma, Houweling, Ferrier, & Ramsey, 2014); however, the Wada test is not perfectly reliable (Janecek et al., 2013; Janecek et al., 2013; Kho et al., 2005; Lanzenberger et al., 2005), and at best provides localization at the level of hemispheric lateralization, rather than specific regions of functional significance. Electrical stimulation mapping is the gold standard for preoperative planning, but correlations with fMRI are imperfect (Giussani et al., 2010). Furthermore, electrical stimulation is restricted to exposed regions of cortex (i.e., surfaces of gyri), unlike the whole-brain coverage afforded by fMRI. Finally, validity might also be assessed by converging results across different tasks. However, it is important to consider the specific demands required of different stimuli and tasks (e.g., words vs. sentences, passive listening vs. repetition, etc.), as these affect the cognitive processes required and thus the brain networks engaged (Peelle, 2012).

Complementing validity is reliability, also an area of increasing interest in neuroimaging research (Botvinik-Nezer et al., 2020; Button et al., 2013; Nosek & Lakens, 2014; Poldrack et al., 2017; Simmons, Nelson, & Simonsohn, 2011). Here we focus on test-retest reliability: the degree to which the same paradigm, run in the same participant, produces the same result. Reliability is independent of validity in the sense that a test could produce inaccurate results but produce the same inaccurate results every time (a watch that always reads 10:10 is reliable, but not very useful). However, reliability is also a crucial part of effective scientific inquiry and clinical application. An additional advantage is that test-retest reliability can be assessed without resorting to additional methods (e.g., Wada tests or cortical stimulation). As a result of more than three decades of functional neuroimaging research on speech and language, as a field there is a reasonably good consensus on some of the major regions and networks we expect to see for different tasks, which can be used to assess face validity. In the current study we thus focus not on identifying new ways to assess validity, but take a retrospective approach to ask: Given the many paradigms used in the field, how often is reliability assessed?

Deciding to focus on reliability is only the first step, as there are a number of possible ways to assess reliability, many of which focus on different aspects of the data (Bennett & Miller, 2010). One broad distinction is between measures that focus on significance compared to those that look at the effect size (i.e., parameter estimates). In the former, overlap measures have been particularly popular. Overlap measures generally binarize result images into significant vs. not-significant labels and then calculate a measure of matching labels across images (Crum et al., 2005); popular approaches include Jaccard and Dice measures. These measures focus on the extent of activation, and philosophically fit reasonably well with standard cluster-based methods of statistical significance. However, they do not tell us anything about the “strength” of the activation (i.e., parameter estimates). For that, the most common

approach is interclass correlation coefficient (ICC) (Noble, Scheinost, & Constable, 2021). The ICC quantifies similarity of two ratings (e.g., test and re-test). Similar to a traditional bivariate correlation, a value of 1.0 indicates near-perfect agreement between values of test and re-test sessions, and a value of 0.0 indicates no agreement between sessions. ICCs can be calculated at the whole brain level, or on summary values from regions of interest. One challenge of ICCs is that there are several different approaches (Shrout & Fleiss, 1979) and it can be challenging to select the correct one to use.

Finally, in the context of language processing, it is also common to look at hemispheric lateralization. Lateralization is commonly quantified using a lateralization index (Wilke & Schmithorst, 2006) although there are various approaches (Bradshaw, Bishop, & Woodhead, 2017). Lateralization has the advantage of being a relatively simple summary measure. At the whole-brain level, lateralization provides no information about spatial specificity of activations (two images could have identical lateralization indices but show activity in completely different regions), although when restricted—for example, to a set of regions of interest—lateralization within those regions can also be calculated.

To characterize the current state of the field with respect to single-subject language localization, we performed a systematic review of fMRI studies reporting single-subject results. Our goals were to document approaches used for assessing the reliability of single-subject results, place language studies in a broader context of fMRI reliability, and, if possible, identify potential design choices associated with improved single-subject reliability.

2. Methods

The information gathered in this systematic review was structured by following the PRISMA guidelines (Liberati et al., 2009), summarized in Fig. 1. Supplemental materials, including data and analysis scripts, are available from https://osf.io/x692b/. We used SunburstR package in R to create the figures (Bostock et al., 2020; R Core Team, 2013).

2.1. Search methods statement

We searched published literature using strategies (Search Strategy in Supplemental materials) designed by a medical librarian for the concepts of functional magnetic resonance imaging (fMRI) and speech or language. Database-supplied limits for English were used. These strategies were established using a combination of controlled vocabulary terms and keywords and were executed in Ovid-Medline, Embase, Scopus, Cochrane Register of Controlled Trials (CENTRAL), and Clinicaltrials.gov. We verified the effectiveness of the search strategies based on how well a set of predefined benchmark papers were captured by the search. All searches were performed on April 8th, 2020. Duplicate citations were removed, leaving 970 unique citations for analysis.

2.2. Additional literature

The initial search did not capture seven previously identified benchmark papers due to missing keywords in the title or abstract. These were added to the final list of papers, which resulted in a total of 977 papers.

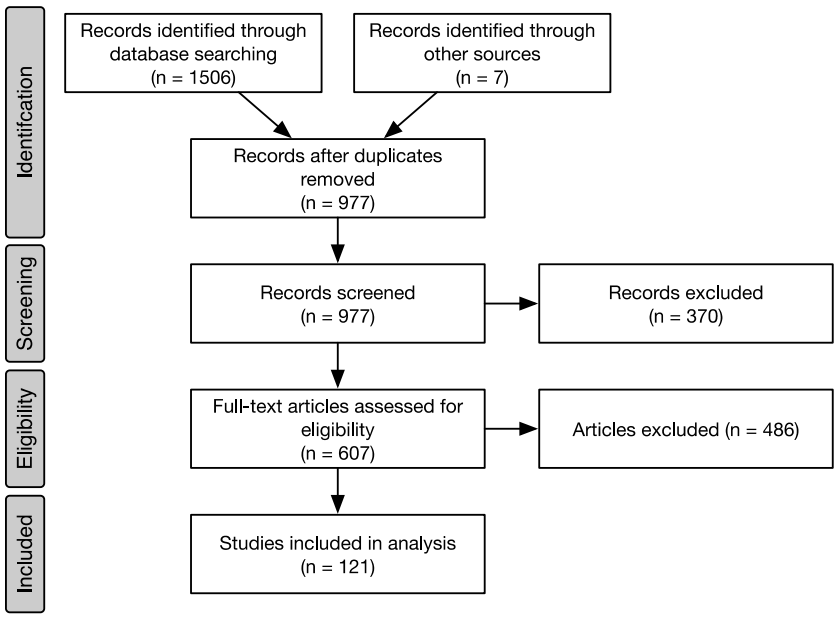

Fig. 1. PRISMA diagram summarizing the literature search.

2.3. Eligibility criteria

We selected all published papers before April 8th, 2020, that met the following criteria.

- Used fMRI as an independent modality or in conjunction with other modalities
- Primary research literature in adults
- Language task used in the experimental design
- Performed single-subject level analysis
- Reported task-based single-subject maps or quantification of single-subject results

Specific search terms are available in supplemental materials. The resulting papers from the database search then underwent abstract screening and full-text screening using the procedure discussed below.

2.4. Abstract screening

We screened abstracts obtained after the literature search to exclude conference abstracts, reviews, technical notes, clinical trials, and other non-primary literature. Articles that were unavailable after a Google Scholar search, duplicate entries, incorrect citations, empty results for abstract contents, studies conducted on children, and abstracts that did not mention fMRI use were also excluded. The extent to which fMRI was used as an imaging modality was not always clear from the abstract alone. Therefore, papers with abstracts that mentioned task-based fMRI, or cited fMRI results, passed the abstract screening step. A group of individuals with an academic background in neuroscience assisted with the abstract level screening. After an initial categorization, abstract eligibility was verified by the first author. All the abstracts were reviewed by at least one person and checked by the first author.

2.5. Full-text screening

We then screened the contents of the articles that passed abstract screening. Papers were excluded from the final list if they only report group-level results and did not present any quantification or visualization of single-subject results. The finalized papers were screened and categorized based on imaging modalities, reliability metrics, language tasks used, and clinical condition of research participants. To the extent possible, we categorized the tasks using labels from the Cognitive Atlas (Poldrack et al., 2011).

2.6. Reliability measures

A common way of quantifying a neuroimaging study's reliability is to assess the test-retest reproducibility. Assuming brain networks have remained stable, performing the same task should result in a similar pattern of brain activity, with differences attributable to measurement error. A concern with reproducibility metrics is the amount of data available to carry out analyses or the technical considerations of repeating an experiment. Nevertheless, including measures for reproducibility may increase confidence in the findings or establish precedents that enhance research practice. The following are the most common measures used by the papers included in this review.

- The *lateralization Index (LI)* is used as a comparative measure of language-related activations between the left and right hemisphere. Although not a reliability measure on its own, a number of papers rely on the reproducibility of LI as a reliability metric (Agarwal et al., 2018; Benjamin et al., 2017; Fernandez et al., 2003; Knecht et al., 2003; Nettekoven, Reck, Goldbrunner, Grefkes, & Weiss Lucas, 2018; Otzenberger, Gounot, Marrer, Namer, & Metz-Lutz, 2005; Voyvodic, 2012; Wilson, Yen, & Eriksson, 2018). The formula for the LI is:

$$LI = \frac{\sum left\ activations - \sum right\ activations}{\sum left\ activations + \sum right\ activations}$$

A score of 1 indicates fully left-lateralized activation, a score of −1 indicates fully right-lateralized activation, and a score of 0 refers to bilateral (i.e., non-lateralized) activation. The LI is of interest in reliability because of the typical dominance of language in the left hemisphere. Lateralization of function may differ in relation to the complexity of the stimulus used in the task design (Bradshaw, Thompson, Wilson, Bishop, & Woodhead, 2017).

- The *activation volume* refers to the volume of activation above some threshold (e.g., in voxels or μL).
- *Sensitivity* refers to the likelihood of a study to identify activity in an expected region of interest (i.e., a true positive rate); *specificity* refers to a lack of activity in regions where none is expected (i.e., a true negative rate).
- *Overlap measures* such as the Dice Coefficient measures the overlap of the number of active voxels across two images (e.g., scanning sessions):

$$Dice\ coefficient = 2\ \frac{Number\ of\ overlapping\ voxels}{Voxels\ in\ first\ session + Voxels\ in\ second\ session}$$

The Dice coefficient for any two images ranges from 0 to 1, with 0 indicating no overlap and 1 indicating complete overlap (Crum, Camara, & Hill, 2006). Several studies included in this review report Dice overlap (or a related overlap measure). One example of a related measure is the Reproducibility Index, as mentioned in Maldjian, Laurienti, Driskill, and Burdette (2002). Here, the metric is obtained by calculating the pairwise ratio of the probability-weighted intersection volume divided by the union volume of surviving activation clusters. Overlap metrics are some of the most intuitive methods of accounting for reproducibility. Factors such as thresholds, cluster sizes, and focus on *a priori* language regions can affect this measure (Wilson, Bautista, Yen, Lauderdale, & Eriksson, 2017).

- *Correlation measures.* The Intra-class Correlation Coefficient (ICC) reflects the relationship of between-subject variance to within-subject variance. Although there are several ways to calculate ICCs (Shrout & Fleiss, 1979), a common approach is to divide the within subject variance by total variance (within subject and between subject). ICC values calculated this way range from 0 to 1, with ICCs near 0 indicating no agreement and ICCs near 1 indicating perfect agreement. It can be used to measure test-retest reliability at a voxel or an ROI for a chosen level of activation. This metric provides a measure of the contribution of individual-level differences in a group result. However, it is vital to treat this measure with caution since it combines information from between-subject and between-sessions variances (i.e., the same ICC value can result from different activation patterns resulting from inadequate models) (for more on ICCs in fMRI, see Chen et al., 2018; Noble et al., 2021). Non-ICC correlation measures include simple bivariate correlations between activation levels over voxels.
- *Activation Distance.* Euclidean Distance (ED) is commonly used to quantify the distance between peak activation across sessions or different task types (Agarwal et al., 2018; Nettekoven et al., 2018; Voyvodic, 2012):

$$ED = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2},$$

where, $x_{1,2}, y_{1,2}, z_{1,2}$ represent coordinates in 3D space. Localization accuracy can be determined based on how close the activation peaks are for subsequent sessions.

3. Results

Our literature search resulted in 977 unique papers, of which 121 met our inclusion criteria of using single-subject level fMRI to

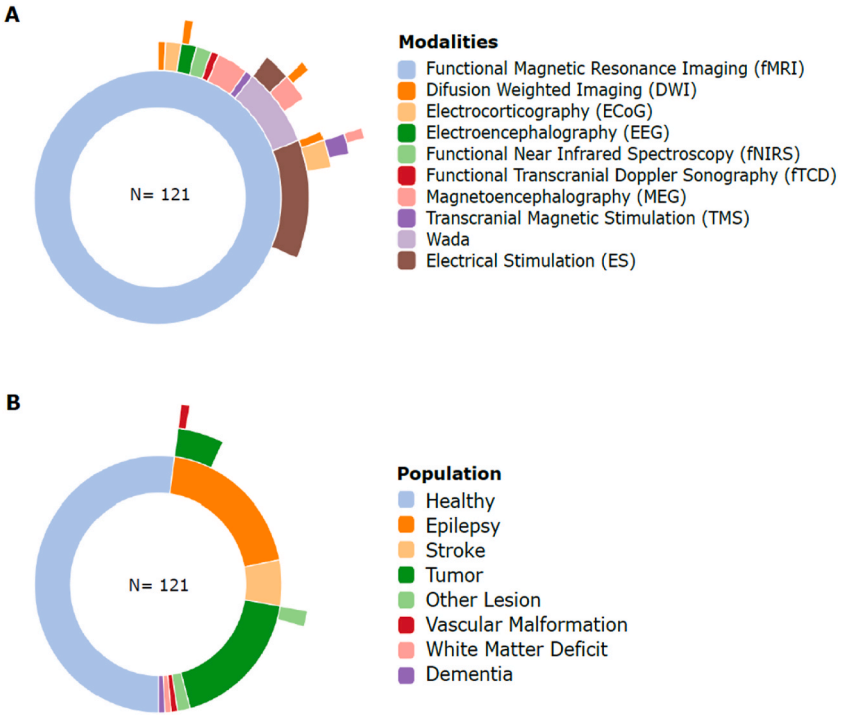

Fig. 2. Characteristics of papers passing screening. Each circle or part of a circle in the Sunburst plot indicates the proportion of studies that belong to the subgroup represented in the legend. **A.** The full inner circle shows the number of studies that used fMRI as one of the imaging modalities. Each subsequent layer represents the proportions of studies with each added modality. **B.** The proportions represent the distribution of population subtypes in the papers. Subsequent layers indicate studies that investigated individuals from more than one population. Details on the number of papers for each category is presented in **Supplemental Information**.

study language processing. Thirty-eight out of the 121 studies used more than one modality to carry out language mapping. The distribution of the modalities used by the papers is presented in Fig. 2A. Language localization studies in a clinical setting are sometimes conducted in the context of pre-surgical planning. Most of the clinical studies that met our inclusion criteria used fMRI as a preliminary method for language localization but relied on the results of invasive mapping methods (such as electrical stimulation) to carry out presurgical planning.

The distribution of the clinical conditions represented in the papers is summarized in Fig. 2B. Out of the 121 included studies, 58 included clinical populations, most of which were multimodal studies. These studies included a wide variety of clinical conditions, epilepsy and tumor being the most common. The study population's distribution highlights the importance of reliable single-subject language mapping methods for both clinical and basic research.

As noted above, we were particularly interested in how many papers reported reliability measures. Of the 121 papers reporting single subject results, 20 reported test-retest sessions and discussed reliability measures as summarized in Table 1. The duration between test and retest sessions in these papers ranged from as close as a few minutes (consecutive test and retest on the same day) to a few years. The most common measures were hemispheric lateralization (16/20), overlap metrics (e.g., Dice) (16/20), and correlation measures (e.g., voxelwise correlation of parameter estimates or Intraclass Correlation Coefficient) (7/20).

Unfortunately, a sensible quantitative comparison of the studies—for example, to determine whether certain paradigms or analyses produce more favorable results—is not possible due to the significant variability in tasks, study population, duration between scans, and thresholding approaches (detailed in Supplementary Table 1). There was a wide range of expressive and receptive tasks used in a multi-task setting (Acioly et al., 2014; Arora et al., 2009; Seghier et al., 2004; Tailby, Abbott, & Jackson, 2017). The most common tasks among these studies were verbal fluency tasks (such as naming and word generation), while less common tasks involved connected speech. The 20 papers that discussed test-retest reliability used block design for their experiments but differed in the technical execution.

Overlap measures were used in 16 of the 20 papers, with 8 using Dice. The 8 Dice papers reported values ranging from 0.34 to 0.66.

Finally, we looked at whether there were changes in the use of test-retest assessments over time. Fig. 3 shows the total number of single-subject fMRI articles we identified, as well as the number reporting test-retest reliability, grouped by year. Although there were more single-subject fMRI papers in later years, the proportion reporting test-retest reliability metrics has stayed fairly constant (1995–2005: 0.21; 2006–2010: 0.16; 2011–2015: 0.09; 2016–2020: 0.18).

4. Discussion

Accurate measurements of regional brain activation in individual participants are essential for clinical research and basic science. In the context of group studies, individual variability in task-based responses has been noted in several domains (Van Horn, Grafton, & Miller, 2008). The focus of our current review was on studies reporting fMRI-based language localization in single subjects. Given the diverse range of language research that spans a broad range of clinical and basic science, our search results might not have captured all the relevant literature. However, the need for reproducible and robust results is essential in any research outcome. The metrics discussed in this paper can also serve as a primer for the most common ways in which reliability metrics can be used while reporting

Table 1
Summary of reliability measures used in the papers that reported test-retest sessions.

Paper	N	Clinical pop.	Lateralization index	Activation volume	Sensitivity/ Specificity	Overlap measure	Correlation measure	Activation Distance
Binder, Rao, et al. (1995)	5	N	X					
Maldjian et al. (2002)	8	N	X			X		
Rutten, Ramsey, van Rijen, and van Veelen (2002)	9	N	X	X		X		
Fernandez et al. (2003)	12	Y	X			X	X	
Knecht et al. (2003)	14	N	X			X		
Otzenberger et al. (2005)	9	N	X			X		
Harrington, Buonocore, and Farias (2006)	10	N	X			X		
Jansen et al. (2006)	10	N	X			X	X	
Chen and Small (2007)	21	Y			X			
Rau et al. (2007)	13	N				X		
Voyvodic (2012)	12	N	X			X		X
Gorgolewski et al. (2013)	10	N				X	X	X
Mahowald and Fedorenko (2016)	32	N	X	X			X	
Benjamin et al. (2017)	22	Y	X			X		
Wilson et al. (2017)	5	N	X		X	X		
Agarwal et al. (2018)	115	Y	X					X
Nettekoven et al. (2018)	16	N	X			X	X	X
Wilson et al. (2018)	30	Y	X	X		X	X	X
Paek, Murray, Newman, and Kim (2019)	16	Y				X	X	
Yen, A, and S (2019)	16	N	X	X	X	X		

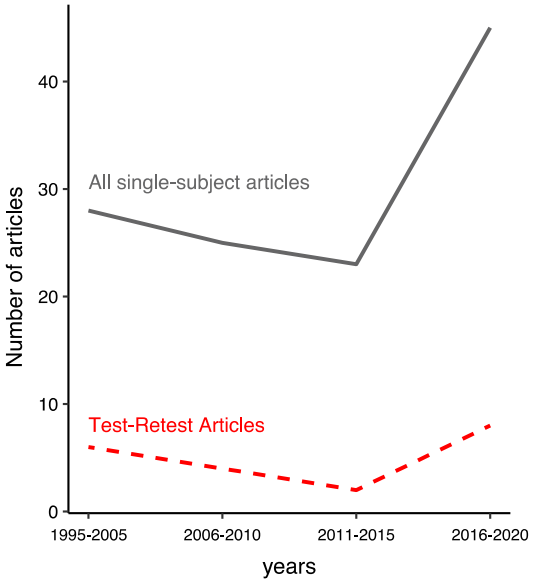

Fig. 3. Count of identified articles across four time ranges. The top line shows all of the 121 articles reporting single-subject data identified by our review; the bottom line the subset that report test-retest reliability measures. [Note the first time bin is larger to accommodate a single paper from 1995; all others occur after 2000.].

neuroimaging results, and we hope to encourage wider adoption of such analyses.

Although test-retest reproducibility is not the only measure of reliability, it is widely used and directly addresses within-subject replicability. Of the papers identified in our search, approximately 1/6 included some measure of test-retest reliability. However, this represents a modest number of studies (20) that vary considerably in the specific language paradigm, participant population, and amount of data collected. Moreover, the choice of metric used to establish reliability was not the same across these studies. Thus, it may prove challenging to generalize existing findings of test-retest reliability to new paradigms or populations. Obtaining an agreement on a standardized approach of quantifying reliability in neuroimaging results would enhance the credibility of research findings. Given that most conclusions are drawn from thresholded statistical maps, the Dice overlap measure is an appealing candidate, although this will be informative with a common threshold across study (e.g., principled control over false positives).

Across a variety of paradigms, we found that average Dice coefficients ranged from 0.34 to 0.66. This range is roughly comparable to that reported by Bennett and Miller (2010), who report a range of average Dice coefficients from 0.23 to 0.79 across a large number of tasks (most not language tasks). As Bennett and Miller highlight, many factors contribute to reliability of fMRI studies, including differences in acquisition, analysis, paradigm, and participants. The number of potential permutations among these factors typically make controlled comparisons impossible. That is, although an individual study may ask "among these four paradigms, which provides the strongest reliability?" (Binder, Swanson, Hammeke, & Sabsevitz, 2008; Wilson et al., 2017), it is far more difficult to answer, "among all the tasks, analysis pipelines, and acquisition parameters available, which provides the strongest reliability?" (which we would all like to know!). However, the variability in reliability measures (such as Dice) suggest that some approaches are more reliable than others. As suggested above, one approach would be to more widely adopt reporting of reliability measures to facilitate optimizing these protocols. More simply, labs might internally use such metrics during task development, to steer them away from tasks that have poor reliability given the idiosyncrasies of their specific questions, acquisition constraints, and analysis pipelines.

In the context of modest test-retest reliability, it is important to note that the issue of single-subject reliability also extends to group-level studies. If the outcome of interest is a group-level univariate map with a relatively large number of participants, inaccuracies in individual participants may have little effect on the result. However, researchers interested in explaining individual differences in brain activation patterns—for example, due to age, language status, hearing loss, etc.—rely on the accuracy of both neural and non-neural estimates of data at an individual level. At a minimum, inaccuracies in measuring brain activity in individual participants will hurt the ability of researchers to detect these effects of interest, or may lead to spurious findings. A related concern applies to multivariate analyses. Even when these analyses are not explicitly designed to localize activity in individual participants, multivariate tests are frequently conducted in single subjects. Error in measuring responses that account for individual differences will likely decrease the accuracy of these analyses.

Another perspective on validity and reliability relates to cross-task comparisons: To the extent that different tasks rely on similar linguistic processes, resulting activation maps should show convergence across task. Although this philosophy of converging evidence across multiple tasks is frequently considered informally, it rarely appears in a quantified analysis. One potential avenue for improving both validity and reliability might be to design studies that explicitly make use of cross-task overlap for robust functional localization.

The high spatial resolution and noninvasiveness of fMRI enable it to complement other language localization approaches to target

potentially more subtle or complex functions than standard clinical practices (Austermuehle et al., 2017; Baciu et al., 2003; Bizzi et al., 2008). However, using fMRI as an independent tool for localizing language for clinical purposes is still in its early stages. Patient and methodological challenges need to be addressed to do so (Beisteiner, Pernet, & Stippich, 2019; Bradshaw, Bishop, & Woodhead, 2017; Bradshaw, Thompson, et al., 2017; Seghier, 2008). An improvement in the reliability of language localization with fMRI might not replace invasive procedures entirely but can help identify subjects that can be assisted without invasive procedures. For those that do undergo invasive procedures, fMRI-based language localization can also assist in monitoring post-surgical recovery.

In theory, it would be beneficial to the field if we were able to identify one or more tasks that meet an accepted standard of reliability to be used confidently in future studies. Unfortunately, reaching such a conclusion is challenged on multiple fronts. First, there is a lack of consensus on the best measure(s) of reliability; and once a measure is chosen, what a reasonable cutoff might be. Second, even with a measure selected, details of implementation matter and vary across the existing literature. For example, for ICCs, are these done at the level of ROIs (atlas-based or functional?) or the whole brain? For overlap measures, what is the threshold (and is it voxelwise, cluster-level, set-level, etc.)? And so on. The amount of variability in the current 20 identified studies is too great to draw any sort of firm conclusion. In addition to further comparisons done within lab (where other factors are held constant) (Wilson et al., 2017), one approach going forward is to converge on reliability metrics. We propose two: Dice overlap and ROI-based ICCs.

For Dice overlap, we propose a threshold of $p < .05$ corrected at the cluster level, using a cluster-forming threshold of $p < .001$ (uncorrected), and calculated using permutation testing (Winkler, Ridgway, Webster, Smith, & Nichols, 2014). We think this is desirable given that many inferences are made at the level of brain regions (rather than individual voxels), and that permutation testing requires fewer assumptions than with random field theory (cf. Nichols & Holmes, 2001; Worsley, 1996).

For ICCs, we propose using group-constrained subject-specific ROIs (Fedorenko et al., 2010; Nieto-Castanon & Fedorenko, 2012). These allow the specific location of an ROI to vary across participants (matching individual functional organization) but within constraints set by group-level results. (Of course, in some cases the group-level constraint may not be appropriate, in which case it should not be used.) It may also be useful to identify atlas-based ROIs that could be kept consistent across study. For example, subdivisions of the IFG based on probabilistic cytoarchitectonic mapping (Amunts et al., 1999) or macroanatomy (Desikan et al., 2006). Of the possible ICC forms, ICC(2, 1) (Shrout & Fleiss, 1979) generally seems most appropriate (Noble et al., 2021).

The combination of Dice overlap and ICCs is appealing because it covers complementary aspects of measurement (binarized statistical significance and continuous effect size). In fact, one approach to benchmarking might be to make use of two metrics, requiring a measure to be over a certain threshold along both dimensions.

Determining the cutoffs for each of these measures is also not a straightforward task. In the context of ICCs, Cicchetti and Sparrow (1981) suggest a threshold of 0.6 for reliability, but a more stringent one of 0.8 for "clinical application" and 0.9 for "individual interpretation". For accurate single-subject functional localization, 0.9 then seems like a reasonable goal. Unfortunately, of the 7 papers we identified as reporting correlation measures, only some reported ICC values, which ranged from ~0 to 0.88 (a single study; semantic decision in Wilson et al., 2018)—none above the 0.9 value. For Dice, anatomical registration methods—including diffeomorphic approaches—historically have average overlap coefficients of about 0.5 (Klein et al., 2009), which is perhaps a reasonable starting point for a threshold. Here we found a maximum Dice of 0.66. It is perhaps not surprising that studies are better at defining "significance" than effect size, as the latter typically requires many more observations (Schönbrodt & Perugini, 2013). It may be that, for task-based fMRI data, more minutes of useable data than typically acquired will be needed to achieve the desired strength of ICC—an observation backed up by a meta-analysis showing a disappointing mean ICC across fMRI studies of 0.397 (Elliott et al., 2020) (although the type of ICCs used in the reviewed studies was not specified).

An important point is that all the studies identified by our search that discussed test-retest reliability carried out univariate analyses. Multivariate analyses are increasingly used to study individual differences (Finn et al., 2015; Woo, Chang, Lindquist, & Wager, 2017), and may well provide reliability that exceeds traditional univariate approaches (Kragel, Han, Kraynak, Gianaros, & Wager, 2021; Noble et al., 2021). With respect to localization, multivariate approaches can vary in their spatial specificity, but searchlight approaches (Etzel, Zacks, & Braver, 2013; Kriegeskorte, Goebel, & Bandettini, 2006) provide an approach for conducting multivariate analyses throughout the brain. Multivariate approaches may therefore prove to be a valuable approach for improving reliability of language localization.

Finally, another avenue that facilitates assessing reliability is to make data freely available. Sharing original data sets would enable researchers to conduct their own measures of reliability across studies. Increasing awareness of data sharing benefits is occurring across scientific disciplines (Poldrack & Gorgolewski, 2014), and publicly available infrastructure for sharing neuroimaging datasets continues to improve (Markiewicz et al., 2021; Poldrack et al., 2013). As illustrated in our findings, a wide variety of tasks, populations, and metrics currently exist, making qualitative and quantitative comparisons across studies challenging.

5. Conclusions

In conclusion, we found that a relatively small number of papers investigating the neurobiology of language—all of which used univariate analysis methods—have assessed the test-retest reliability of single-subject fMRI paradigms. Increased attention to this issue can improve the accuracy and replicability of findings in multiple domains. Some concrete steps towards addressing these concerns could be making reliability metrics such as the Dice coefficient and ICCs a standard part of analyses, reporting both single-subject and group level results to allow transparency, and making data freely available so that researchers can reproduce results or conduct their own reliability analyses.

Funding

Work reported here was supported by the US National Institutes of Health [grant numbers R01 DC019507, R21 DC016086, and T32 EB014855].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Data availability

Data available on OSF

Acknowledgements

We are grateful to Angela Hardi for assistance in conducting the literature search and Stephanie Noble for helpful comments.

Appendix A: Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jneuroling.2023.101163.

References

Acioly, M. A., Gharabaghi, A., Zimmermann, C., Erb, M., Heckl, S., & Tatagiba, M. (2014). Dissociated language functions: A matter of atypical language lateralization or cerebral plasticity? *Journal of Neurological Surgery Part A: Central European Neurosurgery, 75*(1), 64–69. https://doi.org/10.1055/s-0033-1358610

Agarwal, S., Hua, J., Sair, H. I., Gujar, S., Bettegowda, C., Lu, H., et al. (2018). Repeatability of language fMRI lateralization and localization metrics in brain tumor patients. *Human Brain Mapping, 39*(12), 4733–4742. https://doi.org/10.1002/hbm.24318

Amunts, K., Schleicher, A., Burgel, U., Mohlberg, H., Uylings, H. B., & Zilles, K. (1999). Broca's region revisited: Cytoarchitecture and intersubject variability. *Journal of Comparative Neurology, 412*(2), 319–341.

Arora, J., Pugh, K., Westerveld, M., Spencer, S., Spencer, D. D., & Todd Constable, R. (2009). Language lateralization in epilepsy patients: fMRI validated with the Wada procedure. *Epilepsia, 50*(10), 2225–2241. https://doi.org/10.1111/j.1528-1167.2009.02136.x

Austermuehle, A., Cocjin, J., Reynolds, R., Agrawal, S., Sepeta, L., Gaillard, W. D., et al. (2017). Language functional MRI and direct cortical stimulation in epilepsy preoperative planning. *Annals of Neurology, 81*(4), 526–537. https://doi.org/10.1002/ana.24899

Avramescu-Murphy, M., Hattingen, E., Forster, M.-T., Oszvald, A., Anti, S., Frisch, S., et al. (2017). Post-surgical language reorganization occurs in tumors of the dominant and non-dominant hemisphere. *Clinical Neuroradiology, 27*(3), 299–309.

Baciu, M. V., Watson, J. M., McDermott, K. B., Wetzel, R. D., Attarian, H., Moran, C. J., et al. (2003). Functional MRI reveals an interhemispheric dissociation of frontal and temporal language regions in a patient with focal epilepsy. *Epilepsy and Behavior, 4*(6), 776–780. https://doi.org/10.1016/j.yebeh.2003.08.002

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. *Journal of Neurology Neurosurgery and Psychiatry, 85*(5), 581–588. https://doi.org/10.1136/jnnp-2013-305659

Beisteiner, R., Pernet, C., & Stippich, C. (2019). Can we standardize clinical functional neuroimaging procedures? *Frontiers in Neurology, 9*, 1153.

Benjamin, C. F., Walshaw, P. D., Hale, K., Gaillard, W. D., Baxter, L. C., Berl, M. M., et al. (2017). Presurgical language fMRI: Mapping of six critical regions. *Human Brain Mapping, 38*(8), 4239–4255. https://doi.org/10.1002/hbm.23661

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences, 1191*(1), 133–155.

Binder, J. R., Rao, S. M., Hammeke, T. A., Frost, J. A., Bandettini, P. A., Jesmanowicz, A., et al. (1995). Lateralized human brain language systems demonstrated by task subtraction functional magnetic resonance imaging. *Archives of Neurology, 52*(6), 593–601. https://doi.org/10.1001/archneur.1995.00540300067015

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.

Bizzi, A., Blasi, V., Falini, A., Ferroli, P., Cadioli, M., Danesi, U., et al. (2008). Presurgical functional MR imaging of language and motor functions: Validation with intraoperative electrocortical mapping. *Radiology, 248*(2), 579–589.

Bostock, M., Rodden, K., Warne, K., Russell, K., Breitwieser, F., & Yetman, C. (2020). *sunburstR*. CRAN. https://github.com/timelyportfolio/sunburstR.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 1–7.

Bradshaw, A. R., Bishop, D. V., & Woodhead, Z. V. (2017). Methodological considerations in assessment of language lateralisation with fMRI: A systematic review. *PeerJ, 5*, e3557.

Bradshaw, A. R., Thompson, P. A., Wilson, A. C., Bishop, D. V., & Woodhead, Z. V. (2017). Measuring language lateralisation with different language tasks: A systematic review. *PeerJ, 5*, Article e3929.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., et al. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. https://doi.org/10.1038/nrn3475

Chen, E. E., & Small, S. L. (2007). Test-retest reliability in fMRI of language: Group and task effects. *Brain and Language, 102*(2), 176–185. https://doi.org/10.1016/j.bandl.2006.04.015

Chen, G., Taylor, P. A., Haller, S. P., Kircanski, K., Stoddard, J., Pine, D. S., et al. (2018). Intraclass correlation: Improved modeling approaches and applications for neuroimaging. *Human Brain Mapping, 39*(3), 1187–1206. https://doi.org/10.1002/hbm.23909

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86*(2), 127–137. https://www.ncbi.nlm.nih.gov/pubmed/7315877.

Crum, W. R., Camara, O., & Hill, D. L. (2006). Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging, 25*(11), 1451–1461.

Crum, W. R., Camara, O., Rueckert, D., Bhatia, K. K., Jenkinson, M., & Hill, D. L. G. (2005). Generalised overlap measures for assessment of pairwise and groupwise image registration and segmentation. *Medical Image Computing and Computer-Assisted Intervention, 3749*, 99–106.

Dehaene, S., Dupoux, E., Mehler, J., Cohen, L., Paulesu, E., Perani, D., et al. (1997). Anatomical variability in the cortical representation of first and second language. *NeuroReport, 8*(17), 3809–3815.

Desikan, R. S., Segonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage, 31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

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. *Journal of Neuroscience, 40*(23), 4536–4550. https://doi.org/10.1523/JNEUROSCI.2036-19.2020

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., et al. (2020). What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychological Science, 31*(7), 792–806. https://doi.org/10.1177/0956797620916786

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., et al. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods, 16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Etzel, J. A., Zacks, J. M., & Braver, T. S. (2013). Searchlight analysis: Promise, pitfalls, and potential. *NeuroImage, 78*, 261–269. https://doi.org/10.1016/j.neuroimage.2013.03.041

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences, 108*, 16428–16433.

Fedorenko, E., Duncan, J., & Kanwisher, N. G. (2012). Language-selective and domain-general regions lie side by side within Broca's area. *Current Biology, 22*, 2059–2062. https://doi.org/10.1016/j.cub.2012.09.011

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology, 104*, 1177–1194.

Fernandez, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., et al. (2003). Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. *Neurology, 60*(6), 969–975.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., et al. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience, 18*(11), 1664–1671. https://doi.org/10.1038/nn.4135

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*(1), 113–120. https://doi.org/10.1227/01.NEU.0000360392.15450.C9

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., et al. (2017). Precision functional mapping of individual human brains. *Neuron, 95*(4), 791–807 e797. https://doi.org/10.1016/j.neuron.2017.07.011

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

Gould, L., Mickleborough, M. J., Wu, A., Tellez, J., Ekstrand, C., Lorentz, E., et al. (2016). Presurgical language mapping in epilepsy: Using fMRI of reading to identify functional reorganization in a patient with long-standing temporal lobe epilepsy. *Epilepsy Behav Case Rep, 5*, 6–10. https://doi.org/10.1016/j.ebcr.2015.10.003

Gratton, C., Kraus, B. T., Greene, D. J., Gordon, E. M., Laumann, T. O., Nelson, S. M., et al. (2020). Defining individual-specific functional neuroanatomy for precision psychiatry. *Biological Psychiatry, 88*(1), 28–39. https://doi.org/10.1016/j.biopsych.2019.10.026

Hakyemez, B., Erdogan, C., Yildirim, N., Bora, I., Bekar, A., & Parlak, M. (2006). Functional MRI in patients with intracranial lesions near language areas. *The Neuroradiology Journal, 19*(3), 306–312.

Harrington, G. S., Buonocore, M., & Farias, S. T. (2006). Intrasubject reproducibility of functional MR imaging activation in language tasks. *American Journal of Neuroradiology, 27*(4), 938–944.

Janecek, J. K., Swanson, S. J., Sabsevitz, D. S., Hammeke, T. A., Raghavan, M., Mueller, W., et al. (2013). Naming outcome prediction in patients with discordant Wada and fMRI language lateralization. *Epilepsy and Behavior, 27*(2), 399–403. https://doi.org/10.1016/j.yebeh.2013.02.030

Janecek, J. K., Swanson, S. J., Sabsevitz, D. S., Hammeke, T. A., Raghavan, M., Rozman, M. E., et al. (2013). Language lateralization by fMRI and Wada testing in 229 patients with epilepsy: Rates and predictors of discordance. *Epilepsia, 54*(2), 314–322. https://doi.org/10.1111/epi.12068

Jansen, A., Menke, R., Sommer, J., Forster, A. F., Bruchmann, S., Hempleman, J., et al. (2006). The assessment of hemispheric lateralization in functional MRI–robustness and reproducibility. *NeuroImage, 33*(1), 204–217. https://doi.org/10.1016/j.neuroimage.2006.06.019

Jones, M. S., Zhu, Z., Bajracharya, A., Luor, A., & Peelle, J. E. (2022). A multi-dataset evaluation of frame censoring for motion correction in task-based fMRI. *Aperture Neuro, 2*, 1–25. https://doi.org/10.52294/apertureneuro.2022.2.nxor2026

Khateb, A., Martory, M.-D., Annoni, J.-M., Lazeyras, F., de Tribolet, N., Pegna, A. J., et al. (2004). Transient crossed aphasia evidenced by functional brain imagery. *NeuroReport, 15*(5), 785–790.

Kho, K. H., Leijten, F. 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. *Journal of Neurosurgery, 102*(1), 169–173. https://doi.org/10.3171/jns.2005.102.1.0169

Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M.-C., et al. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage, 46*(3), 786–802. https://doi.org/10.1016/j.neuroimage.2008.12.037

Knecht, S., Jansen, A., Frank, A., Van Randenborgh, J., Sommer, J., Kanowski, M., et al. (2003). How atypical is atypical language dominance? *NeuroImage, 18*(4), 917–927.

Kośla, K., Bryszewski, B., Jaskólski, D., Błasiak-Kołacińska, N., Stefańczyk, L., & Majos, A. (2015). Reorganization of language areas in patient with a frontal lobe low grade glioma–fMRI case study. *Polish Journal of Radiology, 80*, 290.

Kragel, P. A., Han, X., Kraynak, T. E., Gianaros, P. J., & Wager, T. D. (2021). Functional MRI Can Be Highly Reliable, but It Depends on What You Measure: A Commentary on Elliott et al. (2020). *Psychological Science*. https://doi.org/10.1177/0956797621989730, 956797621989730.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences, 103*, 3863–3868.

Lanzenberger, R., Wiest, G., Geissler, A., Barth, M., Ringl, H., Wober, C., et al. (2005). FMRI reveals functional cortex in a case of inconclusive Wada testing. *Clinical Neurology and Neurosurgery, 107*(2), 147–151. https://doi.org/10.1016/j.clineuro.2004.06.006

Lee, D., Swanson, S. J., Sabsevitz, D. S., Hammeke, T. A., Scott Winstanley, F., Possing, E. T., et al. (2008). Functional MRI and Wada studies in patients with interhemispheric dissociation of language functions. *Epilepsy and Behavior, 13*(2), 350–356. https://doi.org/10.1016/j.yebeh.2008.04.010

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine, 6*(7), Article e1000100.

Mahowald, K., & Fedorenko, E. (2016). Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability. *NeuroImage, 139*, 74–93. https://doi.org/10.1016/j.neuroimage.2016.05.073

Maldjian, J. A., Laurienti, P. J., Driskill, L., & Burdette, J. H. (2002). Multiple reproducibility indices for evaluation of cognitive functional MR imaging paradigms. *American Journal of Neuroradiology, 23*(6), 1030–1037.

Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The OpenNeuro resource for sharing of neuroscience data. *Elife, 10*. https://doi.org/10.7554/eLife.71774

McGonigle, D. J. (2012). Test–retest reliability in fMRI: or how I learned to stop worrying and love the variability. *NeuroImage, 62*(2), 1116–1120.

McGonigle, D. J., Howseman, A. M., Athwal, B. S., Friston, K. J., Frackowiak, R., & Holmes, A. P. (2000). Variability in fMRI: An examination of intersession differences. *NeuroImage, 11*(6), 708–734.

Nettekoven, C., Reck, N., Goldbrunner, R., Grefkes, C., & Weiss Lucas, C. (2018). Short- and long-term reliability of language fMRI. *NeuroImage, 176*, 215–225. https://doi.org/10.1016/j.neuroimage.2018.04.050

Nichols, T. E., & Holmes, A. P. (2001). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping, 15*, 1–25.

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

Noble, S., Mejia, A. F., Zalesky, A., & Scheinost, D. (2022). Improving power in functional magnetic resonance imaging by moving beyond cluster-level inference. *Proceedings of the National Academy of Sciences of the United States of America, 119*(32), Article e2203020119. https://doi.org/10.1073/pnas.2203020119

Noble, S., Scheinost, D., & Constable, R. T. (2019). A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *NeuroImage, 203*, Article 116157.

Noble, S., Scheinost, D., & Constable, R. T. (2021). A guide to the measurement and interpretation of fMRI test-retest reliability. *Current Opinion in Behavioral Sciences, 40*, 27–32. https://doi.org/10.1016/j.cobeha.2020.12.012

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137–141. https://doi.org/10.1027/1864-9335/a000192

Otzenberger, H., Gounot, D., Marrer, C., Namer, I. J., & Metz-Lutz, M. N. (2005). Reliability of individual functional MRI brain mapping of language. *Neuropsychology, 19*(4), 484–493. https://doi.org/10.1037/0894-4105.19.4.484

Paek, E. J., Murray, L. L., Newman, S. D., & Kim, D. J. (2019). Test-retest reliability in an fMRI study of naming in dementia. *Brain and Language, 191*, 31–45. https://doi.org/10.1016/j.bandl.2019.02.002

Partovi, S., Jacobi, B., Rapps, N., Zipp, L., Karimi, S., Rengier, F., et al. (2012). Clinical standardized fMRI reveals altered language lateralization in patients with brain tumor. *American Journal of Neuroradiology, 33*(11), 2151–2157.

Peelle, J. E. (2012). The hemispheric lateralization of speech processing depends on what "speech" is: A hierarchical perspective. *Frontiers in Human Neuroscience, 6*, 309.

Polczynska, M., Benjamin, C. F., Japardi, K., Frew, A., & Bookheimer, S. Y. (2016). Language system organization in a quadrilingual with a brain tumor: Implications for understanding of the language network. *Neuropsychologia, 86*, 167–175. https://doi.org/10.1016/j.neuropsychologia.2016.04.030

Polczynska, M., Japardi, K., & Bookheimer, S. Y. (2017). Lateralizing language function with pre-operative functional magnetic resonance imaging in early proficient bilingual patients. *Brain and Language, 170*, 1–11. https://doi.org/10.1016/j.bandl.2017.03.002

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafo, M. R., et al. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience, 18*(2), 115–126. https://doi.org/10.1038/nrn.2016.167

Poldrack, R., Barch, D. M., Mitchell, J., Wager, T., Wagner, A. D., Devlin, J. T., et al. (2013). Toward open sharing of task-based fMRI data: The OpenfMRI project. *Frontiers in Neuroinformatics, 7*, 12.

Poldrack, R. A., & Gorgolewski, K. J. (2014). Making big data open: Data sharing in neuroimaging. *Nature Neuroscience, 17*(11), 1510–1517. https://doi.org/10.1038/nn.3818

Poldrack, R., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., et al. (2011). The cognitive atlas: Toward a knowledge foundation for cognitive neuroscience [original research]. *Frontiers in Neuroinformatics, 5*(17). https://doi.org/10.3389/fninf.2011.00017

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. *Journal of Neurosurgery, 97*(1), 21–32.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria.

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. *American Journal of Neuroradiology, 28*(7), 1346–1353. https://doi.org/10.3174/ajnr.A0581

Röder, B., Stock, O., Bien, S., Neville, H., & Rösler, F. (2002). Speech processing activates visual cortex in congenitally blind humans. *European Journal of Neuroscience, 16*(5), 930–936.

Roland, J. L., Hacker, C. D., Breshears, J. D., Gaona, C. M., Hogan, R. E., Burton, H., et al. (2013). Brain mapping in a patient with congenital blindness - a case for multimodal approaches. *Frontiers in Human Neuroscience, 7*, 431. https://doi.org/10.3389/fnhum.2013.00431

Ruff, I. M., Petrovich Brennan, N. M., Peck, K. K., Hou, B. L., Tabar, V., Brennan, C. W., et al. (2008). Assessment of the language laterality index in patients with brain tumor using functional MR imaging: Effects of thresholding, task selection, and prior surgery. *AJNR American Journal of Neuroradiology, 29*(3), 528–535. https://doi.org/10.3174/ajnr.A0841

Rutten, G. J., Ramsey, N. F., van Rijen, P. C., & van Veelen, C. W. (2002). Reproducibility of fMRI-determined language lateralization in individual subjects. *Brain and Language, 80*(3), 421–437. https://doi.org/10.1006/brln.2001.2600

Schönbrodt, F., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*, 609–612.

Seghier, M. L. (2008). Laterality index in functional MRI: Methodological issues. *Magnetic Resonance Imaging, 26*(5), 594–601.

Seghier, M. L., Lazeyras, F., Pegna, A. J., Annoni, J. M., Zimine, I., Mayer, E., et al. (2004). Variability of fMRI activation during a phonological and semantic language task in healthy subjects. *Human Brain Mapping, 23*(3), 140–155. https://doi.org/10.1002/hbm.20053

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037//0033-2909.86.2.420

Sierpowska, J., Gabarros, A., Ripolles, P., Juncadella, M., Castaner, S., Camins, A., et al. (2013). Intraoperative electrical stimulation of language switching in two bilingual patients. *Neuropsychologia, 51*(13), 2882–2892. https://doi.org/10.1016/j.neuropsychologia.2013.09.003

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Smith, S. M., Beckmann, C. F., Ramnani, N., Woolrich, M. W., Bannister, P. R., Jenkinson, M., et al. (2005). Variability in fMRI: A re-examination of inter-session differences. *Human Brain Mapping, 24*(3), 248–257.

Tailby, C., Abbott, D. F., & Jackson, G. D. (2017). The diminishing dominance of the dominant hemisphere: Language fMRI in focal epilepsy. *NeuroImage: Clinical, 14*, 141–150.

Tomasino, B., Marin, D., Canderan, C., Maieron, M., Budai, R., Fabbro, F., et al. (2014). Involuntary switching into the native language induced by electrocortical stimulation of the superior temporal gyrus: A multimodal mapping study. *Neuropsychologia, 62*, 87–100. https://doi.org/10.1016/j.neuropsychologia.2014.07.011

Van Horn, J. D., Grafton, S. T., & Miller, M. B. (2008). Individual variability in brain activity: A nuisance or an opportunity? *Brain Imaging and Behavior, 2*(4), 327.

Voyvodic, J. T. (2012). Reproducibility of single-subject fMRI language mapping with AMPLE normalization. *Journal of Magnetic Resonance Imaging, 36*(3), 569–580.

Wilke, M., & Schmithorst, V. J. (2006). A combined bootstrap/histogram analysis approach for computing a lateralization index from neuroimaging data. *NeuroImage, 33*(2), 522–530. https://doi.org/10.1016/j.neuroimage.2006.07.010

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

Wilson, S. M., Yen, M., & Eriksson, D. K. (2018). An adaptive semantic matching paradigm for reliable and valid language mapping in individuals with aphasia. *Human Brain Mapping, 39*(8), 3285–3307. https://doi.org/10.1002/hbm.24077

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage, 92*, 381–397. https://doi.org/10.1016/j.neuroimage.2014.01.060

Woo, C. W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience, 20*(3), 365–377. https://doi.org/10.1038/nn.4478

Worsley, K. J. (1996). The geometry of random images. *Chance, 9*, 27–39.

Yen, M. D., A, T. W., & S, M. (2019). Adaptive paradigms for mapping phonological regions in individual participants [Article]. *NeuroImage, 189*, 368–379. https://doi.org/10.1016/j.neuroimage.2019.01.040

Zhang, J., Anderson, J. R., Liang, L., Pulapura, S. K., Gatewood, L., Rottenberg, D. A., et al. (2009). Evaluation and optimization of fMRI single-subject processing pipelines with NPAIRS and second-level CVA. *Magnetic Resonance Imaging, 27*(2), 264–278. https://doi.org/10.1016/j.mri.2008.05.021