RESEARCH ARTICLE

WILEY

An adaptive semantic matching paradigm for reliable and valid language mapping in individuals with aphasia

Stephen M. Wilson[1] (ORCID) | Melodie Yen[1] (ORCID) | Dana K. Eriksson[2]

[1]Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, Nashville, Tennessee

[2]Department of Speech, Language, and Hearing Sciences, University of Arizona, Tucson, Arizona

Correspondence
Stephen M. Wilson, Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, 1215 21st Ave S, MCE 8310, Nashville, TN 37232.
Email: stephen.m.wilson@vanderbilt.edu

Funding information
National Institutes of Health (National Institute on Deafness and Other Communication Disorders), Grant/Award Numbers: R01 DC013270 and R21 DC016080

Abstract

Research on neuroplasticity in recovery from aphasia depends on the ability to identify language areas of the brain in individuals with aphasia. However, tasks commonly used to engage language processing in people with aphasia, such as narrative comprehension and picture naming, are limited in terms of reliability (test–retest reproducibility) and validity (identification of language regions, and not other regions). On the other hand, paradigms such as semantic decision that are effective in identifying language regions in people without aphasia can be prohibitively challenging for people with aphasia. This paper describes a new semantic matching paradigm that uses an adaptive staircase procedure to present individuals with stimuli that are challenging yet within their competence, so that language processing can be fully engaged in people with and without language impairments. The feasibility, reliability and validity of the adaptive semantic matching paradigm were investigated in sixteen individuals with chronic post-stroke aphasia and fourteen neurologically normal participants, in comparison to narrative comprehension and picture naming paradigms. All participants succeeded in learning and performing the semantic paradigm. Test-retest reproducibility of the semantic paradigm in people with aphasia was good (Dice coefficient = 0.66), and was superior to the other two paradigms. The semantic paradigm revealed known features of typical language organization (lateralization; frontal and temporal regions) more consistently in neurologically normal individuals than the other two paradigms, constituting evidence for validity. In sum, the adaptive semantic matching paradigm is a feasible, reliable and valid method for mapping language regions in people with aphasia.

KEYWORDS
aphasia, language mapping, reliability, validity

1 | INTRODUCTION

Damage to brain regions involved in language processing typically results in aphasia. However, language function can improve over time, either spontaneously (Kertesz & McCabe, 1977; Swinburn, Porter, & Howard, 2004), or potentially in response to behavioral (Brady, Kelly, Godwin, Enderby, & Campbell, 2016), neuromodulatory (Shah, Szaflarski, Allendorfer, & Hamilton, 2013) or pharmacological (Berthier & Davila, 2014) interventions. Recovery from aphasia after damage to language regions of the brain is thought to depend on neural plasticity, that is, functional reorganization of surviving brain regions such that they take on new or expanded roles in language processing (Geranmayeh, Brownsett, & Wise, 2014; Heiss & Thiel, 2006; Nadeau, 2014; Price & Crinion, 2005; Saur et al., 2006; Saur & Hartwigsen,

2012; Turkeltaub, Messing, Norise, & Hamilton, 2011). There is currently a great deal of interest in characterizing the nature of this putative process of functional reorganization. A better understanding of when and how reorganization takes place, how different patterns of reorganization depend on patient-specific factors, and how different patterns are associated with better or worse outcomes, could inform the design of new therapies, and could facilitate optimal targeting of specific interventions to individual patients (Fridriksson, Richardson, Fillmore, & Cai, 2012; Thompson & den Ouden, 2008).

This line of research critically depends on being able to identify brain regions involved in language processing in individual patients, and being able to determine with statistical rigor whether they change over time (Kiran et al., 2013; Meinzer et al., 2013; Wilson, Bautista, Yen, Lauderdale, & Eriksson, 2017). Language areas of the brain can be

identified with functional magnetic resonance imaging (fMRI) using language mapping paradigms, which generally contrast conditions that involve language processing to conditions that do not (Binder, Swanson, Hammeke, & Sabsevitz, 2008). To support research on functional reorganization of language regions in recovery from aphasia, a language mapping paradigm needs to meet at least three criteria.

First, it must be feasible and appropriate for individuals with aphasia. The most common clinical application of language mapping is in presurgical contexts, in which the aim is to avoid resecting eloquent cortex. These patients usually do not have significant language impairments, so they can readily perform a range of language tasks (Binder et al., 2008). In contrast, individuals with aphasia are impaired in language processing, which implies that they will likely experience difficulty with language tasks, and depending on the task, may not be able to perform it at all. It is difficult to interpret activation maps associated with failure to perform a task (Price, Crinion, & Friston, 2006), presenting a challenge: how can language processing be engaged in a controlled manner in people whose language function is by definition compromised?

Second, the language mapping paradigm must be reliable. In other words, a map of language regions obtained in a given participant on a given occasion should be reproducible in the same participant on a different occasion (Bennett & Miller, 2010). This is referred to as test-retest reproducibility. Research on neuroplasticity requires being able to distinguish genuine changes from scan-to-scan variability (Kiran et al., 2013; Meinzer et al., 2013; Wilson et al., 2017).

Third, the language mapping paradigm must be valid, that is, it must identify all and only the regions that are actually critical for language, as opposed to perceptual, motor, cognitive, and executive regions that may be recruited by different tasks. This implied dichotomy between language and non-language regions is an oversimplification, given that domain-general regions are also needed to support language processing (Fedorenko & Thompson-Schill, 2014), but in most neurologically normal individuals, there are core frontal and temporal language regions, which are lateralized to the left hemisphere (Knecht et al., 2003; Seghier, Kherif, Josse, & Price, 2011; Springer et al., 1999; Tzourio-Mazoyer et al., 2010; Bradshaw, Thompson, Wilson, Bishop, & Woodhead, 2017). In neurological populations, other patterns of organization may be observed (Berl et al., 2014).

We will show with reference to three commonly used paradigms that these three criteria—feasibility, reliability and validity—have rarely if ever been simultaneously met in research to date. Narrative comprehension paradigms have often been used in aphasia recovery research (Crinion, Warburton, Lambon Ralph, Howard, & Wise, 2006; Crinion & Price, 2005; Warren, Crinion, Lambon Ralph, & Wise, 2009). Narrative comprehension is feasible for most individuals with aphasia, since it requires no responses, and comprehension and/or recall can be quantified in post-scan testing. Acoustically matched control conditions, such as backwards speech, have typically been used. Probably due to the relatively unconstrained nature of both the language and the control conditions, the reliability of narrative comprehension paradigms has been empirically demonstrated to be moderate at best (Harrington, Buonocore, & Farias, 2006; Maldjian, Laurienti, Driskill, & Burdette, 2002;

Wilson et al., 2017). The validity of narrative comprehension for mapping the language network is somewhat marginal. While activations are somewhat left-lateralized (Harrington et al., 2006; Maldjian et al., 2002; Wilson et al., 2017), there is a substantial right hemisphere component in neurologically normal individuals (Crinion et al., 2003), so that right hemisphere activation in patients cannot be interpreted as reflecting reorganization (Crinion & Price, 2005). Moreover, only temporal lobe regions are activated with good sensitivity (Crinion, Lambon Ralph, Warburton, Howard, & Wise, 2003; Wilson et al., 2017), so frontal regions important for language processing are not amenable to study.

Picture naming is another simple and feasible task that has been widely used (Abel, Weiller, Huber, Willmes, & Specht, 2015; Fridriksson, Baker, & Moser, 2009; Postman-Caucheteux et al., 2010). Anomia is ubiquitous in aphasia, and correct and incorrect items can be separated and compared (Fridriksson et al., 2009; Postman-Caucheteux et al., 2010). Scrambled pictures have often been used as control stimuli. Picture naming paradigms are moderately reliable, but the reproducible activations tend to be in bilateral sensorimotor areas which are uninformative with respect to language localization (Harrington et al., 2006; Jansen et al., 2006; Meltzer, Postman-Caucheteux, McArdle, & Braun, 2009; Rau et al., 2007; Rutten, Ramsey, van Rijen, & van Veelen, 2002; Wilson et al., 2017). Validity is significantly limited: picture naming paradigms show only modest lateralization, and often do not activate frontal and/or temporal sites (Harrington et al., 2006; Jansen et al., 2006; Rau et al., 2007; Rutten et al., 2002; Wilson et al., 2017).

Semantic decision paradigms are widely used clinically in presurgical language mapping with non-aphasic patients. For example, Binder et al. (1997) described a paradigm that has been used in many subsequent studies (Binder et al., 2008). In the language condition, participants hear a series of animal names and have to decide if each animal is found in the United States and used by humans. In the control condition, participants make perceptual decisions about sequences of high and low tones (deciding whether exactly two high tones occurred). Semantic decision paradigms show good test-retest reproducibility (Fernández et al., 2003; Fesl et al., 2010), probably due to the highly constrained processing involved in both the language and control tasks. Moreover, the deep, active, challenging language processing engendered by semantic tasks reliably activates strongly lateralized frontal and temporal language regions in people without language deficits, providing evidence for validity (Binder et al., 1997, 2008; Fesl et al., 2010; Janecek et al., 2013; Springer et al., 1999; Szaflarski et al., 2008).

However, despite the good reliability and validity of semantic decision paradigms, the feasibility of these tasks in individuals with aphasia is questionable. Variants of the Binder task have been used in aphasia recovery research (Eaton et al., 2008; Griffis et al., 2017; Kim, Karunanayaka, Privitera, Holland, & Szaflarski, 2011; Szaflarski, Allendorfer, Banks, Vannest, & Holland, 2013). Not surprisingly, given the complexity of the tasks, patient performance has been very poor. For instance, in one study, individuals with unresolved aphasia post middle cerebral artery stroke performed at chance not only on the semantic condition (47.6%) but also on the tone decision control condition (52.2%; Szaflarski et al., 2013). It appears to be prohibitively challenging for many individuals with aphasia to maintain the complex verbal instructions

pertaining to each condition, switch between them, apply the criteria to evaluate incoming stimuli, and select between two different response buttons.

In sum, these three language mapping paradigms that have often been used in studies investigating neuroplasticity in recovery from aphasia all have significant limitations in terms of feasibility, validity and/or reliability. This has been a roadblock to progress in understanding functional reorganization of language regions in recovery from aphasia. Various other paradigms have also been used in studies of people with aphasia, including word repetition (Heiss, Kessler, Thiel, Ghaemi, & Karbe, 1999; Weiller et al., 1995), sentence comprehension (Saur et al., 2006), verb generation (Allendorfer, Kissela, Holland, & Szaflarski, 2012; Weiller et al., 1995), and simpler semantic decision paradigms that certain patients are able to perform (Kiran et al., 2015; Robson et al., 2014; Sharp, Scott, & Wise, 2004; van Oers et al., 2010; Zahn et al., 2004). These paradigms vary widely in terms of their feasibility (Price et al., 2006), reliability (Wilson et al., 2017) and validity (Bradshaw et al., 2017), but to our knowledge, no studies to date have evaluated all three of these psychometric properties of the paradigms they have used.

To address this limitation, we developed a paradigm that builds on the strong reliability and validity of semantic decision paradigms, yet is modified so as to be suitable for individuals with aphasia. Specifically, we developed an adaptive semantic matching paradigm in which a conceptually simple task is melded with an adaptive staircase procedure (Leek, 2001) that tailors item difficulty to individual performance, so that the same paradigm can be used to map language regions in people with different degrees of language impairment, as well as in individuals with normal language function. The adaptive nature of the paradigm entails that regardless of their level of language function, all people are performing a focused, challenging task that is highly constrained in terms of the linguistic and other processing required.

We investigated the feasibility, reliability and validity of this adaptive semantic paradigm, in comparison to narrative comprehension and picture naming paradigms. Sixteen individuals with chronic and stable post-stroke aphasia were each scanned on two separate occasions, and fourteen neurologically normal individuals were each scanned once. During each scanning session, all participants performed the three language mapping paradigms. Structural imaging data were also acquired, and language deficits were quantified. Feasibility was evaluated in terms of the ability of individuals with aphasia to learn the tasks and perform them in the scanner. Reliability was assessed in individuals with aphasia by quantifying the similarity of the activation maps obtained in the two sessions using the Dice coefficient of similarity. Validity was evaluated primarily in the neurologically normal individuals, in whom there are strong *a priori* expectations that lateralized frontal and temporal language regions should be activated; validity in individuals with aphasia was also examined in an exploratory manner. Finally, we examined whether our findings were impacted by several analysis parameters: region of interest, absolute or relative vowelwise thresholds, and cluster extent threshold.

2 | METHODS

2.1 | Adaptive semantic matching paradigm

The adaptive semantic matching paradigm comprises two conditions: a semantic matching task, and a perceptual matching task. The tasks are presented in alternating 20-s blocks in a simple AB block design. There are 10 blocks per condition, for a total scan time of 400 s (6:40). Each block contains between 4 and 10 items (inter-trial interval 5–2 s), depending on the level of difficulty.

In the semantic condition, each item consists of a pair of words, which are presented one above the other in the center of the screen (Figure 1a). Half of the pairs are semantically related (e.g., *boy-girl*, *lizard-snake*, *grass-lawnmower*), while the other half are not related (e.g., *walnut-bicycle*). The participant presses a single button with a finger of their left hand if they decide that the words are semantically related. If the words are not related, they do nothing.

In the perceptual condition, each item comprises a pair of false font strings, presented one above the other (Figure 1b). Half of the pairs are identical (e.g., ΔΘδЋϟ-ΔΘδЋϟ), while the other half are not identical (e.g., ΔΘδЋϟ-ϟΔɦκΔ). The participant presses the button if the strings are identical, and does nothing if they differ.

The semantic and perceptual tasks are equivalent in terms of sensorimotor, executive and decision-making components, yet make differential demands on language processing. Task-switching demands are minimized because both conditions involve an essentially similar task: pressing a button to matching pairs. The use of just a single button obviates the need to learn an arbitrary association between "match" and one button, and "mismatch" and another button. The left hand is used for the button press because many individuals with post-stroke aphasia have right-sided hemiparesis.

Critically, both the semantic task and the perceptual task are independently adaptive to participant performance. Each task has seven levels of difficulty. Whenever the participant makes two successive correct responses on a given condition, they move to the next highest level of difficulty on the subsequent trial of that condition. Whenever they make an incorrect response, they move two levels down on the next trial. This is a 2-down-1-up adaptive staircase with weighted step sizes (up twice as large as down), which theoretically should converge at just over 80% accuracy (García-Pérez, 1998). Note that the difficulty level is manipulated independently for the semantic and perceptual conditions, even though sets of items from the two conditions are interleaved due to the AB block design.

Similar contrasts between semantic matching and perceptual matching tasks, without any adaptive component, have been used in several previous studies. Most similar to the present study, Seghier et al. (2004) contrasted a semantic category matching task with a false font string matching perceptual task. Several other studies have contrasted synonym judgments with match–mismatch judgments on letter strings (e.g., Fernández et al., 2001; Gitelman, Nobre, Sonty, Parrish, & Mesulam, 2005).

2.1.1 | Manipulation of item difficulty

Two versions of the experiment were constructed, differing in terms of how item difficulty was manipulated. Most of the data reported in this
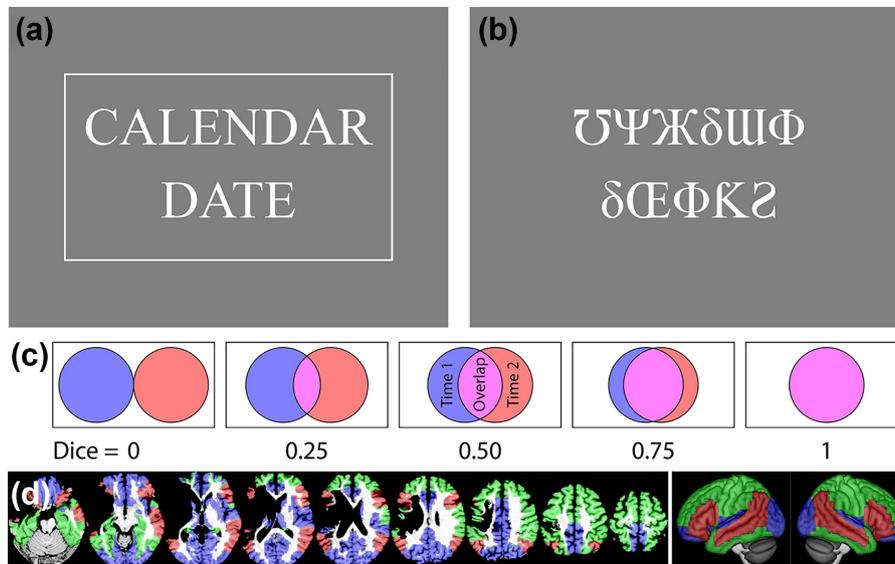
FIGURE 1 Methodological details. (a) Example semantic item. This item is a match, and is shown surrounded by a box that appears when the "match" button is pressed (the box confirms the button press, but no information on accuracy is provided). (b) Example perceptual item. This item is a mismatch, so the button should not be pressed. (c) An illustration of how the Dice coefficient of similarity captures the extent of overlap between two thresholded images. (d) Regions of interest in a representative participant and projected onto the lateral surfaces of a template brain. ROI 1 (*Brain*) encompassed the whole brain. ROI 2 (*Supra*) encompassed regions shown in red, green or blue. ROI 3 (*Lang+*) corresponded to regions shown in red or green. ROI 4 (*Lang*) is shown in red

paper were acquired using the first version. However some ceiling effects were observed in neurologically normal individuals, so a revised version was constructed, which we advocate the use of in future studies. Here both versions are described. Example items from both conditions at each level are shown for the original version in Table 1 and for the revised version in Table 2.

In the original version, the difficulty of semantic items was manipulated by varying four factors: as the level of difficulty increased, words had lower lexical frequency, words were less concrete, pairs of matching words were less closely related to one another, and the presentation rate was faster. Word pairs were extracted from the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 1998). This is a large dataset of word pairs derived from a procedure in which participants were presented with single words and asked to "provide the first word that came to mind that is meaningfully related or strongly associated to the presented word" (Nelson et al., 1998). A MATLAB program was used to select a total of 1327 pairs that varied systematically according to the first three factors listed above. Lexical frequency was obtained from the American National Corpus (Reppen, Ide, & Suderman, 2005), concreteness norms were

TABLE 1 Original version of the adaptive semantic matching paradigm

Level	Frequency	Concreteness	Forward strength	Backward strength	Match example	Mismatch example	Perceptual match	Perceptual mismatch
1	7.38 ± 1.01	601 ± 52	0.64 ± 0.11	0.29 ± 0.27	*girl boy*	*king mom*	ΘΘδ ΘΘδ	ΣϽΣ ΚΞΘΣΘΓ
2	6.51 ± 1.49	547 ± 94	0.45 ± 0.12	0.16 ± 0.17	*nest bird*	*fuel tree*	ΔδΦ℧ ΔδΦ℧	ЖᵭΣϽδΘ ƧΘΘΣ
3	6.13 ± 1.36	548 ± 100	0.32 ± 0.09	0.10 ± 0.13	*calendar date*	*orange Wet*	Ͻ℧ϽδΔ Ͻ℧ϽδΔ	℧ΨЖδɯΦ δŒΦΚƧ
4	5.67 ± 1.28	530 ± 108	0.26 ± 0.07	0.07 ± 0.10	*onion cry*	*shiny ballet*	ΘΘδΓΓ ΘΘδΓΓ	ᵭΣϽδΘ ΘƧΘΘΣ
5	5.32 ± 1.29	503 ± 112	0.21 ± 0.05	0.05 ± 0.07	*organize neat*	*limb owe*	ᵭΔδΦ℧ ᵭΔδΦ℧	५ΓϽ५ΨŒ ƧΓϽ५ŒϽ
6	4.52 ± 1.33	467 ± 116	0.17 ± 0.05	0.04 ± 0.06	*magnify enlarge*	*bristle sour*	ɯ५ΞΞΦϽ ɯ५ΞΞΦϽ	ΨɯΘЖΚΞ ΨΚΘŒΕΚΞ
7	3.42 ± 1.33	391 ± 108	0.10 ± 0.05	0.02 ± 0.04	*menthol eucalyptus*	*caress astonish*	Ж५ϽΞЈℏ Ж५ϽΞЈℏ	ΔΘЖᵭΘŒ ΔΘᵭΘŒ

TABLE 2 Revised version of the adaptive semantic matching paradigm

Level	Frequency	Concreteness	Length	Age of acquisition	Match example	Mismatch example	Perceptual match	Perceptual mismatch
1	7.85 ± 1.20	544 ± 91	4.16 ± 0.89	4.11 ± 0.93	*rat mouse*	*read house*	ꝶꙊꙄẟ ꝶꙊꙄẟ	Ꙋꙁꚅẟꙍꝶ ẟŒꝶꝂꝂ
2	6.98 ± 1.39	525 ± 99	4.68 ± 1.28	5.22 ± 1.37	*kiss love*	*sea fork*	ꝊꙊẟꝪꝪ ꝊꙊẟꝪꝪ	ꝬꙄꙍꝊ ꝊꙌꝊꝊꙍ
3	6.17 ± 1.42	518 ± 96	5.20 ± 1.43	6.22 ± 1.60	*salt vinegar*	*shark empty*	ꝬꙂẟꝊꙊ ꝬꙂẟꝊꙊ	ꝂꝪꝬꝂꝊŒꙄ ꙌꝪꝬꝂŒꙄꙄ
4	5.78 ± 1.64	483 ± 104	5.80 ± 1.69	7.26 ± 1.83	*symbol ornament*	*squirrel selection*	ŒꝊꙄꙄꝽ ŒꝊꙄꙄꝽ	ẟꝊꝂꝊꙎ ꙄꝂꝊꝊꙎ
5	5.35 ± 1.59	462 ± 106	6.37 ± 1.96	8.45 ± 1.94	*limousine carriage*	*fever honeycomb*	ẟꙄꙊꙊꝂꝊꝊ ẟꙄꙊꙊꝂꝊꝊ	ẟꙄꝽꝪꝽ ꝽꙄꝽꝪꝽ
6	4.88 ± 1.67	419 ± 112	6.95 ± 2.20	9.54 ± 2.02	*arrangement agreement*	*ceremony expulsion*	ꙍꝽꝪꝪꝊꙄ ꙍꝽꝪꝪꝊꙄ	ꝊꙍꝊꝀꝂꝊ ꝊꝀꝊꝊꝀꙍꝂ
7	3.93 ± 1.54	367 ± 103	8.06 ± 2.22	11.46 ± 1.97	*catastrophe upheaval*	*intermission socialist*	ꝀꝽꙄꝪꙄꝽꝽ ꝀꝽꙄꝪꙄꝽꝽ	ẟꝊꝀꝊŒ ẟꝊẟꝊŒ

obtained from the MRC database (Coltheart, 1981), and degree of relatedness between pairs of words was defined in terms of forward strength (cue-to-target strength) and backward strength (target-to-cue strength) of the free association norms (Nelson et al., 1998).

The paradigm was revised with the main goal of making the difficult levels more difficult, in order to avoid ceiling effects in people without language impairments. In the revised version of the paradigm, the difficulty of semantic items was manipulated by lexical frequency, concreteness, degree of relatedness, and presentation rate, similar to the original version, but also by word length and age of acquisition. The free association norms were not used. Instead, all words with concreteness ratings in the MRC database (∼4,000) were retrieved. A difficulty metric was computed for each of these words, by summing the *z* scores for frequency (Reppen et al., 2005), concreteness (Coltheart, 1981), age of acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), and length (number of letters); frequency and age of acquisition were doubly weighted. The pairwise semantic distances between all of the 4,000 words were estimated with *snaut* (Mandera, Keuleers, & Brysbaert, 2017), a prediction-based model of distributional semantics derived from corpora. For each word, the ten most related words output by *snaut* were considered as possible match items (these items were not necessarily actually closely related to the target word, given the limitations of the computational model). A matching word was manually selected where the salience of the semantic relationship was subjectively similar to the difficulty metrics of the words involved. For instance, the word *rat* is frequent, acquired early, concrete, and short, so it was paired with a closely related associate, *mouse* to create an easy item. In contrast, *elopement* is infrequent, acquired late, abstract, and long, so it was paired with the tangentially related *consummation* (rather than closer possible associates such as *marriage*) to create a difficult item. A total of 1,934 items were created in this way, then ordered by the average difficulty metric of the pair of words.

For both versions of the experiment, mismatching items were created by shuffling adjacent pairs in the difficulty-ordered list, then manually adjusting any incidentally created matches. In both versions, item difficulty in the perceptual condition was manipulated in two ways: as the level of difficulty increased, mismatching pairs were more similar, and presentation rate was faster. However in the revised version of the experiment, mismatching items were more similar at the lower levels, making the perceptual task more difficult at these levels.

In order to match sensorimotor and executive demands across the semantic and perceptual conditions, it is necessary to yoke presentation rate across conditions. Presentation rate is adjusted at the start of each semantic block and remains fixed for the upcoming semantic block and perceptual block. The "ideal" inter-trial interval for each condition is defined as the block length (20 s) divided by the ideal number of items per block (4–10, for difficulty levels 1–7). The number of items per block is then selected to be as large as possible without exceeding the average of the two "ideal" inter-trial intervals.

2.1.2 | Training

There are three phases of training. In the first phase, which typically takes about five minutes, the examiner explains the tasks to the participant in language that is appropriate for the individual, taking into account the nature and severity of their aphasia, if any. It is recommended that the examiner be experienced in communicating with individuals with aphasia. As the examiner explains the tasks, they present items manually using key presses. Match and mismatch semantic and perceptual items can be presented at any level of difficulty and with any timing appropriate to the situation. The examiner can also press the "match" button to demonstrate, and then have the participant practice pressing it. The examiner can present as many practice items as are necessary for the participant to learn the task.

Once the participant is comfortable responding to individually presented items, the second training phase begins. In this phase, stimuli are delivered continuously in a block design, identical to the functional

imaging experiment except that the presentation rate is not yoked across conditions (because it does not need to be). The participant thus becomes familiar with the structure and presentation rate of the actual experiment. The researcher can stop this training phase when the participant is familiar with the paradigm. For most individuals with aphasia, two or three blocks of each condition should be presented (about two minutes). The third phase of training is the same as the second phase, except that it takes place after the participant has been placed in the scanner, for instance, during acquisition of localizer or structural images. In this way, the participant becomes accustomed to performing the task in the unfamiliar environment of the scanner bore. One to two minutes are generally sufficient, but this third phase of training does not add to overall testing time, since it takes place concurrent with acquisition of structural images.

2.1.3 | Technical implementation

The semantic paradigm is implemented in a MATLAB program called *AdaptiveLanguageMapping* using the Psychophysics Toolbox version 3 (Brainard, 1997; Pelli, 1997), which requires MATLAB R2012a or later. The *AdaptiveLanguageMapping* program has been tested on Linux, Windows and Mac OS X, and has no dependencies besides the freely available Psychophysics Toolbox. This program can also be used to present the other two paradigms described in this paper (narrative comprehension and picture naming). A manual is provided describing the operation of the program. Log files are generated allowing for analysis of behavioral data. *AdaptiveLanguageMapping* is freely available for download at http://aphasialab.org/alm.

2.2 | Narrative comprehension paradigm

The narrative comprehension paradigm was also a simple AB block design with two conditions: narrative comprehension, and backwards speech. As in the semantic paradigm, there were 10 blocks per condition, for a total scan time of 400 s (6:40). In the narrative blocks, the opening pages of an audiobook were presented, in one session *Who was Albert Einstein?* (Brallier, 2002) and in the other session *Who were the Beatles?* (Edgers, 2006), each of which were written for children aged 8–12 years, and so contain relatively simple language. In the backwards speech blocks, the same segments were played in reverse. To avoid sentences being split across blocks, the blocks varied in length from 13.65 to 26.59 s (mean = 19.63 ± 3.45 s). There was a gap of 0.37 s between blocks.

To aid comprehension by providing a visual reminder of the topic of the narrative, an iconic picture of Albert Einstein or the Beatles was displayed in the center of the screen prior to the start of the run and throughout the run (during both conditions). After the end of the scanning session, individuals with aphasia were asked six true/false questions about the narrative they had heard.

2.3 | Picture naming paradigm

The picture naming paradigm was a jittered rapid event-related design with two conditions: real pictures and scrambled pictures. An event-related design was used in order to match previous applications of

similar paradigms in aphasia, in which the goal has often been to separately model correct and incorrect items. Continuous acquisition rather than sparse sampling was used because the delay of the hemodynamic response ensured that speech-related and neural effects were temporally distinct (Birn, Bandettini, Cox, & Shaker, 1999), and the former were largely accounted for in preprocessing (see below). There were 40 real pictures and 20 scrambled pictures, presented in a total scanning time of 400 s (6:40). Each picture was displayed for 3 s. Participants were instructed to name the real pictures, overtly if possible, and to just look at the scrambled pictures. The mean inter-trial interval (from the onset of one item to the onset of the next item) was 6.5 ± 2.2 s (range 4–14 s), leaving adequate time for spoken responses. In between trials, participants were asked to fixate on a central crosshair.

Spoken responses were recorded with a scanner-compatible microphone (FORMI-III, OptoAcoustics, Mazor, Israel). Responses were coded as correct, incorrect, or no response. Reaction times were measured from the presentation time to the onset of the first response, not including any fillers, false starts or fragments.

The stimuli were colorized versions (Rossion & Pourtois, 2004) of the Snodgrass and Vanderwart (1980) pictures. Only items with mono-morphemic targets and name agreement of at least 60% were used. The mean length of the target names was 3.9 ± 0.9 phonemes (range 3–6), the mean log frequency of the targets based on the HAL corpus (Lund & Burgess, 1996) was 8.8 ± 1.5 (range 4.9–13.2), and the mean name agreement was 89.8% ± 9.8% (range 62%–100%). The means and distributions of these variables were matched across two versions of the paradigm that were presented in the two sessions.

2.4 | Participants

Sixteen individuals with chronic post-stroke aphasia were recruited from an aphasia center in Tucson, Arizona, or were prior participants in aphasia research at the University of Arizona. The inclusion criteria were (a) persistent and stable aphasia of any etiology; (b) at least six months post stroke; (c) aged 18–90; (d) fluent and literate in English premorbidly. The exclusion criteria were (a) dementia; (b) major psychiatric disorders; (c) serious substance abuse. Of the 16 participants, 15 had experienced left hemisphere strokes, while one (A5) had experienced bilateral strokes, with the right hemisphere stroke being more extensive.

Fourteen neurologically normal participants were recruited mostly from a neighborhood listserv in Tucson, Arizona. They reported no neurological or psychiatric history. The Mini Mental State Examination (Folstein, Folstein, & McHugh, 1975) was administered to each participant, and scores ranged from 27 to 30.

Demographic information is presented in Table 3. All but four participants were native speakers of English. Two individuals with aphasia and two neurologically normal participants were native speakers of Spanish whose primary language was now English and who were fluent in English.

TABLE 3 Demographic and language data

	Individuals with aphasia	Neurologically normal participants
Number of participants	16	14
Age (years)	60.4 ± 14.8 (32–79)	53.1 ± 15.1 (27–80)
Sex (M/F)	12/4	8/6
Handedness (R/L)	14/2	10/4
Education (years)	15.1 ± 2.5 (12–20)	17.1 ± 1.9 (14–20)
Days post stroke	1,955 ± 1,220 (208–3960)	
Quick aphasia battery		
Spoken word comprehension	9.5 ± 1.2 (5.3–10.0)	10.0 ± 0.0 (10.0–10.0)
Sentence comprehension	5.0 ± 2.9 (0.7–9.9)	9.3 ± 1.3 (5.4–10.0)
Word finding	4.8 ± 2.8 (0.0–9.4)	9.8 ± 0.5 (8.5–10.0)
Grammatical construction	5.2 ± 3.4 (0.0–9.7)	9.8 ± 0.2 (9.5–10.0)
Speech motor programming	6.6 ± 4.3 (0.0–10.0)	10.0 ± 0.0 (10.0–10.0)
Repetition	5.5 ± 2.7 (0.0–9.9)	9.7 ± 0.4 (8.8–10.0)
Written word comprehension	9.7 ± 1.0 (5.8–10.0)	
Reading aloud	5.5 ± 3.3 (0.0–9.9)	9.6 ± 0.4 (8.8–10.0)
Quick aphasia battery overall	6.0 ± 2.3 (1.4–9.7)	9.8 ± 0.3 (9.0–10.0)
Other language measures		
WAB Aphasia quotient	67.8 ± 24.5 (18.9–98.8)	
CAT Written word to picture (/30)	27.4 ± 2.8 (20–30)	
Pyramids and Palm Trees (/14)	13.9 ± 0.3 (13–14)	

CAT = Comprehensive Aphasia Test; WAB = Western Aphasia Battery.

For the revised version of the adaptive semantic matching paradigm, a second group of 16 neurologically normal participants were recruited from a neighborhood listserv in Nashville, Tennessee (age 57.0 ± 15.0 years (range 23–77 years); 6 male, 10 female; 12 right-handed, 3 left-handed, 1 ambidextrous; education 16.7 ± 2.2 years (range 12–20 years); MMSE range 27–30). These participants did not complete the other two paradigms, and were scanned on a different scanner, so they were not directly compared to the other participants.

All participants gave written informed consent and were modestly compensated for their time. The study was approved by the institutional review boards at the University of Arizona and Vanderbilt University, and all study procedures were performed in accordance with Declaration of Helsinki.

2.5 | Language assessments

Individuals with aphasia completed three study sessions. In the first session, the study was explained to them and they provided written informed consent, demographic and medical history information, and were screened for MRI safety. To characterize language deficits, they completed one of three equivalent forms of Quick Aphasia Battery (QAB; Wilson, Eriksson, Schneck, & Lucanie, 2018), as well as the Western Aphasia Battery (WAB; Kertesz, 1982). Because the adaptive semantic matching task depends on comprehension and semantic processing of written words, patients then completed the written word to picture matching subtest of the Comprehensive Aphasia Test (Swinburn et al., 2004), written word to picture matching using the word

comprehension items from another form of the QAB, and a 14-item short version (Breining et al., 2015) of the Pyramids and Palm Trees Test (Howard & Patterson, 1992). In the second and third sessions, the other two forms of the QAB were administered, along with written word to picture matching using QAB items.

Language data are shown in Table 3. Four patients presented with Broca's aphasia per clinical impression (A1, A2, A3, and A4), one patient was non-verbal with good comprehension (A5), five patients had conduction aphasia (A6, A7, A8, A9, and A10), one patient was agrammatic in production but with good comprehension, fitting no traditional subtype (A11), four patients had anomic aphasia (A12, A13, A14, and A15), and one patient was almost completely recovered (A16). The patients spanned a range of aphasia severity: per WAB criteria, five had severe aphasia, four had moderate aphasia, six had mild aphasia, and one was within normal limits. The patients' language function is described in more detail elsewhere (Wilson et al., 2018). Written word comprehension was excellent in all but one of the patients, the only exception being the person with the most severe Broca's aphasia, who comprehended about two thirds of written words. Non-verbal semantic function was intact in all patients.

Neurologically normal individuals completed just one study session, in which their language function was evaluated with a single form of the QAB.

2.6 | Neuroimaging

Individuals with aphasia were scanned with structural and functional MRI during their second and third study sessions. The first imaging

session took place 15.9 ± 3.8 days after the consent/behavioral session, and there were 11.4 ± 6.8 days between the two imaging sessions. Neurologically normal participants were scanned during their only study session.

Prior to entering the scanner, each participant was trained on the three language mapping tasks. MRI data were acquired on a Siemens Skyra 3 Tesla scanner with a 20-channel head coil at the University of Arizona. Visual stimuli were presented on a 24″ MRI-compatible LCD monitor (BOLDscreen, Cambridge Research Systems, Rochester, UK) positioned at the end of the bore, which participants viewed through a mirror mounted to the head coil. Auditory stimuli were presented using insert earphones (S14, Sensimetrics, Malden, MA) padded with foam to attenuate scanner noise and reduce head movement. The presentation volume was adjusted to a comfortable level for each participant.

For each of the three language mapping paradigms, T2*-weighted BOLD echo planar images were collected with the following parameters: 200 volumes + 3 initial volumes discarded; 30 axial slices in interleaved order; slice thickness = 3.5 mm with a 0.9 mm gap; field of view = 240 × 218 mm; matrix = 86 × 78; repetition time (TR) = 2,000 ms; echo time (TE) = 30 ms; flip angle = 90°; voxel size = 2.8 × 2.8 × 4.4 mm. The order of the three language mapping paradigms was counterbalanced across participants. High resolution T2-weighted images were acquired coplanar with the functional images in each session, to aid coregistration.

For anatomical reference and lesion delineation, T1-weighted MPRAGE structural images (voxel size = 0.9 × 0.9 × 0.9 mm) were acquired (in the first imaging session in patients). In patients only, to provide more information for lesion delineation, T2-weighted FLAIR images (voxel size = 0.5 × 0.5 × 2.0 mm) were acquired in the second imaging session.

The second group of neurologically normal participants were scanned on a Philips Achieva 3 Tesla scanner with a 32-channel head coil at Vanderbilt University. Visual stimuli were projected onto a screen at the end of the bore, which participants viewed through a mirror mounted to the head coil. T2*-weighted BOLD echo planar images were collected with the following parameters: 200 volumes + 4 initial volumes discarded; 35 axial slices in interleaved order; slice thickness = 3.0 mm with 0.5 mm gap; field of view = 220 × 220 mm; matrix = 96 × 96; repetition time (TR) = 2,000 ms; echo time (TE) = 30 ms; flip angle = 75°; SENSE factor = 2; voxel size = 2.3 × 2.3 × 3.5 mm. Coplanar T2-weighted images and T1-weighted structural images were also acquired.

2.7 | Analysis of neuroimaging data

The functional data were first preprocessed with tools from AFNI (Cox, 1996). Head motion was corrected, with six translation and rotation parameters saved for use as covariates. Next, the data were detrended with a Legendre polynomial of degree 2, and smoothed with a Gaussian kernel (FWHM = 6 mm). Then, independent component analysis (ICA) was performed using the FSL tool *melodic* (Beckmann & Smith, 2004). Noise components were manually identified with reference to the criteria of Kelly et al. (2010) and removed using *fsl_regfilt*.

First level models were fit independently for each of the six functional runs (two sessions, three paradigms per session). The adaptive semantic matching and narrative comprehension paradigms were modeled with simple boxcar functions. For the picture naming paradigm, explanatory variables were created for picture items and scrambled items; an additional analysis was performed in which correct and incorrect items were modeled separately.

Task models were convolved with a hemodynamic response function (HRF) based on the difference of two gamma density functions (time to first peak = 5.4 s, FWHM = 5.2 s; time to second peak = 15 s; FWHM = 10 s; coefficient of second gamma density = 0.09), and fit to the data with the program *fmrilm* from the FMRISTAT package (Worsley et al., 2002). The six motion parameters were included as covariates, as were time-series from white matter and CSF regions (means of voxels segmented as white matter or CSF in the vicinity of the lateral ventricles) to account for nonspecific global fluctuations, and three cubic spline temporal trends.

Lesions were manually demarcated based on T1-weighted and FLAIR images. The T1-weighted anatomical images were warped to MNI space using unified segmentation in SPM5 (Ashburner & Friston, 2005) with cost-function masking of the lesion (Brett, Leff, Rorden, & Ashburner, 2001). Functional images were coregistered with structural images via coplanar T2-weighted structural images using SPM, and warped to MNI space. Functional images were inclusively masked with a gray matter mask obtained by smoothing the segmented gray matter proportion image with a 4 mm FWHM Gaussian kernel, then applying a cutoff of 0.25. Patient images were exclusively masked with the lesion mask. These two steps were performed in order to increase reliability by excluding activations that are likely to be spurious.

To identify brain regions activated by each paradigm when normal language function is intact, a random effects analysis was carried out on the 14 neurologically normal individuals. Each contrast was thresholded at voxelwise $p < .005$, then corrected for multiple comparisons at $p < .01$ based on cluster extent according to Gaussian random field theory as implemented in SPM5 (Worsley et al., 1996). This stringent corrected threshold was chosen because Gaussian random field theory can inflate false positive rates (Eklund, Nichols, & Knutsson, 2016). A group analysis was performed in the same way on the second group of 16 neurologically normal participants who were scanned on the revised version of the semantic paradigm.

2.8 | Quantification of reliability

Reliability of fMRI paradigms is generally assessed by scanning the same participants two or more times, and then calculating a similarity metric between the activations obtained on each occasion (Bennett & Miller, 2010). Our similarity metric was the Dice coefficient of similarity (Rombouts et al., 1997), which is a measure of the extent of overlap between thresholded activation maps. The Dice coefficient is calculated as 2. $V_{overlap}/(V_1 + V_2)$ where $V_{overlap}$ is the number of overlapping voxels, V_1 is the number of voxels activated in the first scan, and V_2 is the number of voxels activated in the second scan (Figure 1c). The advantages of the Dice coefficient are that it is easy to interpret

(Bennett & Miller, 2010), is widely used (e.g., Fernández et al., 2003; Fesl et al., 2010; Gross & Binder, 2014; Harrington et al., 2006; Rutten et al., 2002), can be calculated in any individual without reference to a group (Bennett & Miller, 2010); and yields a single metric of overall activation similarity encompassing all brain regions under consideration (Wilson et al., 2017). In these last two respects, the Dice coefficient is more useful than the intraclass correlation coefficient, another metric sometimes used in research on language mapping paradigms (Fernández et al., 2003; Eaton et al., 2008; Meltzer et al., 2009). In this paper, Dice coefficients will be described as poor (<0.40), fair (0.40–0.60), good (0.60–0.75) or excellent (≥0.75), following Cicchetti (1994), since Dice coefficients are conceptually related to the kappa statistic (Zijdenbos, Dawant, Margolin, & Palmer, 1994).

2.9 | Quantification of validity

Validity is the extent to which a language mapping paradigm identifies all and only the regions that are actually critical for language processing. The validity of language mapping paradigms has been investigated in several different ways. Concurrent validity has been examined by comparing fMRI to the Wada test (intracarotid amobarbital test) for language lateralization (Janecek et al., 2013; Woermann et al., 2003), and to electrocortical stimulation mapping for language localization within a hemisphere (Giussani et al., 2010). These invasive approaches are not feasible in our population; moreover, neither are infallible as a ground truth: fMRI has been shown to out-perform Wada (Janecek et al., 2013), and stimulation mapping is limited to the exposed surfaces of gyri, and language areas are not identified at all in a significant minority of patients (Sanai, Mirzadeh, & Berger, 2008). Therefore, we took a different approach to quantifying validity.

Specifically, it is firmly established that language is lateralized to the left hemisphere in the vast majority of neurologically normal individuals (Knecht et al., 2003; Seghier et al., 2011; Springer et al., 1999; Tzourio-Mazoyer et al., 2010; Bradshaw et al., 2017). This is why aphasia results from left hemisphere lesions, and not from right hemisphere lesions. Therefore, a valid language mapping paradigm should yield left-lateralized activation maps in the majority of neurologically normal participants. Lateralization indices (LIs) were calculated according to the standard formula: $LI = (V_{Left} - V_{Right})/(V_{Left} + V_{Right})$, where V_{Left} is the number of voxels activated in the left hemisphere, and V_{Right} is the number of voxels activated in the right hemisphere. LI ranges from -1 (all activation in the right hemisphere) to $+1$ (all activation in the left hemisphere). In individuals with aphasia, who were each scanned twice, the LI was averaged across the two sessions. When no voxels were activated in either hemisphere, LI was set to 0.

Another known fact about language organization is that frontal and temporal regions in the dominant hemisphere are involved in language processing in the majority of neurologically normal individuals (Knecht et al., 2003; Seghier et al., 2011; Springer et al., 1999; Tzourio-Mazoyer et al., 2010; Bradshaw et al., 2017). Therefore, a second assay of validity was the sensitivity of paradigms to detect expected language activations in these regions. The frontal ROI was defined as the inferior frontal gyrus (AAL regions 11, 13 and 15; Tzourio-Mazoyer et al.,

2002), while the temporal ROI was defined as the middle temporal gyrus (85), angular gyrus (65), and the ventral part of the superior temporal gyrus (81); specifically, voxels within 8 mm of the middle temporal gyrus. The extent of activation within these ROIs was compared.

Validity was assessed primarily in the neurologically normal group, since only in these participants were there clear expectations about the lateralization and localization of language regions. However, in exploratory analyses, lateralization and sensitivity to frontal and temporal language regions were also calculated for individuals with aphasia, even though lateralization may have changed, and frontal and/or temporal language regions in some cases have been destroyed.

2.10 | Analysis parameter sets

Calculations of reliability and validity are strongly influenced by analysis parameters such as region of interest (ROI), vowelwise threshold, and cluster extent threshold (Wilke & Lidzba, 2007; Wilson et al., 2017). The impact of these three parameters was systematically explored using different combinations that will be referred to as analysis parameter sets. However, an *a priori* parameter set was also selected based on previous findings as explained below, for display of individual activation maps and for statistical comparisons between the three paradigms.

Four regions of interest were defined, each of which was symmetrical across the two hemispheres (Figure 1d). The first ROI, labeled *Brain*, was simply the whole brain, i.e. no mask was applied except for the gray matter and lesion masks described above. LIs were not calculated for the *Brain* ROI, since the inclusion of the cerebellum would make LI invalid in this case, given that language activations in the cerebellum are usually, but not always, contralateral to cortical activations (Gelinas, Fitzpatrick, Kim, & Bjornson, 2014).

The second ROI, labeled *Supra* (Figure 1d, red, green or blue), consisted of all supratentorial gray matter regions (AAL regions 1–90), that is cortical gray matter, the thalamus and the basal ganglia.

The third ROI, termed *Lang+* (Figure 1d, red or green), consisted of a very liberal set of brain regions which are either known language regions or plausible candidate regions for functional reorganization. Included in this ROI were the inferior frontal gyrus (AAL regions 11/13/15 and their right hemisphere counterparts); middle frontal gyrus (7); superior frontal gyrus (3/23); supplementary motor area (19); precentral gyrus (1); postcentral gyrus (57); supramarginal gyrus (63); angular gyrus (65); the remainder of the inferior parietal lobule (61); the superior parietal lobule (59); the ventral part of the superior temporal gyrus (81), specifically voxels within 8 mm of the middle temporal gyrus; the middle temporal gyrus (85); the temporal pole (83/87); the inferior temporal gyrus (89); the fusiform gyrus (55); the parahippocampal gyrus (39); and the hippocampus (37). This ROI would be a good choice for studies of functional reorganization in aphasia, because it can be expected to improve reliability by excluding potentially spurious activations in unlikely loci for functional reorganization. Accordingly, this ROI was selected as the *a priori* ROI for this study.

The fourth ROI, termed *Lang* (Figure 1d, red), was narrowly defined as likely language regions and their right hemisphere homotopic counterparts, specifically: the inferior frontal gyrus (11/13/15); the ventral
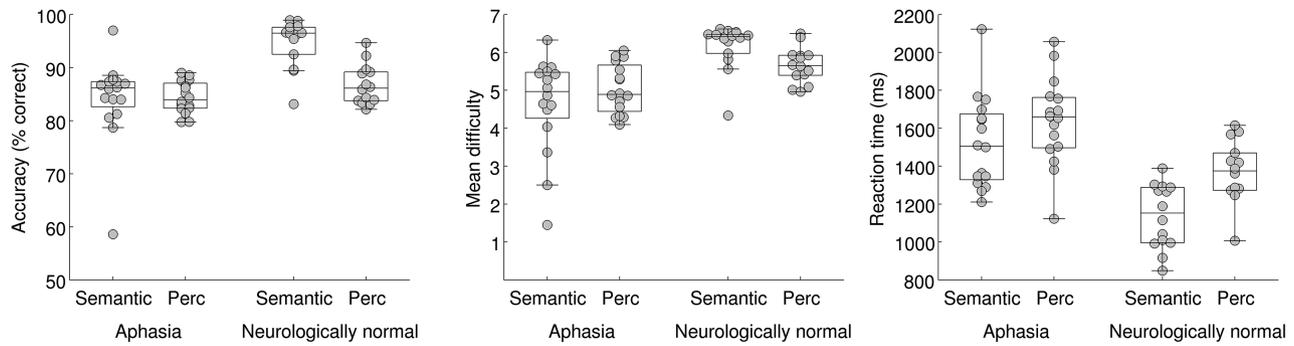
FIGURE 2 Behavioral data for the adaptive semantic matching paradigm. Accuracy, item difficulty, and reaction time on the semantic and perceptual control conditions in individuals with aphasia and neurologically normal participants. Perc = Perceptual

part of the superior temporal gyrus (81) defined as above; the middle temporal gyrus (85); and the angular gyrus (65). This ROI would be a good choice for studies interested primarily in language lateralization within known language regions.

Images of t statistics were thresholded at 14 different voxelwise thresholds. Seven of these were absolute: $p < .1$, $p < .05$, $p < .01$, $p < .005$, $p < .001$, $p < .0005$, and $p < .0001$. The other seven thresholds were relative: the top 10%, 7.5%, 5%, 4%, 3%, 2%, or 1% of most highly activated voxels were considered active, the denominator being the total number of voxels included in the *Brain* ROI (so that the total extent of activation would be consistent across ROIs). Note that relative thresholds of 10%, 7.5%, and 5% were not calculated for the *Lang* ROI, since the *Lang* ROI included only 13.7% ± 1.1% of the voxels in the Brain ROI, so these more liberal relative thresholds would have entailed over 30% of voxels in the ROI being considered active. Approaches involving relative thresholds often improve reliability (Gross & Binder, 2014; Knecht et al., 2003; Voyvodic, 2012; Wilson et al., 2017), therefore the *a priori* threshold was a relative threshold, specifically top 5%, similar to Gross and Binder (2014).

Four cluster extent thresholds were applied: none, 1, 2, and 4 cm³. The *a priori* cluster extent threshold was 2 cm³, selected based on previous findings for a narrative comprehension paradigm (Wilson et al., 2017).

3 | RESULTS

3.1 | Behavioral data

For the adaptive semantic matching paradigm, overall accuracy was plotted as a function of condition (semantic, perceptual) in individuals with aphasia and neurologically normal participants (Figure 2). A mixed effects ANOVA revealed a significant interaction of group by condition ($F(1, 28) = 7.84$; $p = .0091$), such that patients performed equivalently accurately on the two conditions (semantic: 84.1 ± 7.9 (SD)%; perceptual: 84.4% ± 3.0%; $|t(15)| = 0.13$, $p = .90$), while neurologically normal participants were more accurate on the semantic condition (94.7% ± 4.5%) than the perceptual condition (86.7% ± 3.7%; $|t(13)| = 4.91$; $p = .0003$). The better performance of the neurologically normal group on the semantic condition was presumably due to the items not being difficult enough even at the most difficult level. To

address this limitation, the semantic paradigm was subsequently revised; data from the revised paradigm are presented below in the subsection entitled "Revised adaptive semantic stimuli".

The mean difficulty level of items presented was plotted as a function of condition in individuals with aphasia and neurologically normal participants (Figure 2). A mixed effects ANOVA revealed a significant interaction of group by condition ($F(1, 28) = 4.46$; $p = .044$), such that for patients, mean item difficulty level did not differ between the semantic condition (4.64 ± 1.27) and the perceptual condition (5.04 ± 0.65; $|t(15)| = 1.11$; $p = .29$), while neurologically normal participants performed at a higher level of difficulty on the semantic condition (6.16 ± 0.61) than on the perceptual condition (5.66 ± 0.47; $|t(13)| = 2.61$; $p = .022$). This significant interaction was expected due to language deficits in individuals with aphasia, but note that the precise pattern of differences is less important, because while both tasks have seven levels of difficulty, the degree of difficulty of the two tasks is not inherently matched at each level.

Reaction times on correct items were plotted as a function of condition in individuals with aphasia and neurologically normal participants (Figure 2). There was no significant interaction of group by condition ($F(1, 28) = 1.12$; $p = 0.30$). There was a main effect of group, such that patients responded slower than neurologically normal participants ($F(1, 28) = 24.23$; $p < .0001$), and a main effect of condition, such that reaction times were faster in the semantic condition ($F(1, 28) = 15.22$; $p = .0005$). While it would be preferable for reaction times to be equivalent, it is less problematic for responses to be faster in the condition of interest than the control condition, as opposed to vice versa, which could result in areas modulated by time on task being misidentified as language areas (Binder, Medler, Desai, Conant, & Liebenthal, 2005).

For the narrative comprehension paradigm, comprehension was assessed after each scanning session with six true/false questions. Individuals with aphasia answered these questions with a mean accuracy of 90.6% ± 11.7% (range 66.7%–100%), indicating that narratives were generally attended and comprehended, albeit imperfectly in some cases. Neurologically normal participants were not asked the questions, but all reported having heard and attended to the narratives.

For the picture naming paradigm, individuals with aphasia responded correctly to 65.2% ± 33.7% of items (range 0%–100%), provided incorrect responses to 21.3% ± 24.8% of items (range 0%–

80.0%), and did not respond to 13.5% ± 24.1% of items (range 0%–100%). The mean reaction time on correct items was 1,504 ± 250 ms (range 1,045–2,074 ms). There was one non-verbal participant who did not provide any overt responses to any item, and another participant who provided only one correct response across the two sessions. To avoid excluding these two individuals, we report functional imaging analyses carried out on all items, not just correct items. Analyses based on correct items only were also performed (excluding these two participants) and yielded essentially similar activation patterns and estimates of reliability and validity (data not shown).

Neurologically normal participants were more accurate (98.8% ± 1.3% range 97.5%–100%) than individuals with aphasia ($|t(15.05)| = 3.99$; $p = .0012$) on the picture naming task, and responded more quickly (1,153 ± 155 ms; range 862–1,379 ms; $|t(21.71)| = 4.46$; $p = .0002$).

3.2 | Language activation maps

To identify the brain regions activated by each paradigm in the absence of any damage to language networks, random effects group analyses were carried out for each paradigm in the group of 14 neurologically normal participants.

For the adaptive semantic matching paradigm, the semantic condition was contrasted to the perceptual condition (Figure 3a; Table 4). This contrast activated the left inferior frontal gyrus (pars opercularis, triangularis and orbitalis) extending to the precentral sulcus and posterior middle frontal gyrus, and also into the temporal pole; the left posterior superior temporal sulcus and middle temporal gyrus, extending into the angular gyrus, anteriorly along the superior temporal sulcus, and ventrally into the inferior temporal gyrus and fusiform gyrus; the anterior hippocampi bilaterally; and the right cerebellum.

For the narrative comprehension paradigm, the narrative comprehension condition was contrasted to the backwards speech condition (Figure 4a). The most prominent activations were in the left and right temporal lobes. In the left hemisphere, there was activation along the length of the left superior temporal sulcus and adjacent superior temporal and middle temporal gyri, extending from the angular gyrus to the temporal pole. In the right hemisphere, a similar pattern of activation was seen except that activation was absent in the posterior superior temporal sulcus. There was also an activation in the left inferior frontal gyrus extending to the precentral sulcus. Other areas activated were the head of the left caudate nucleus; the precuneus spanning the midline; the superior frontal gyrus spanning the midline; and the right cerebellum.

For the picture naming paradigm, picture naming was contrasted to viewing scrambled pictures (Figure 5a). This contrast activated an extensive set of bilateral regions including ventral occipito-temporal cortex, the posterior superior temporal gyrus, the precentral and postcentral gyri, inferior frontal gyri, and numerous other cortical and subcortical regions. With the exception of inferior frontal activity, which was somewhat left-lateralized, the pattern of activation was generally symmetrical, and largely reflected sensorimotor processes (object perception, speech motor, hearing one's own voice).

3.3 | Test–retest reproducibility

Each patient's two activation maps derived from the adaptive semantic matching paradigm in the two separate sessions are shown in Figure 3b. Activation maps for a subset of four patients (P9–P12, the third row of Figure 3b) are shown for the narrative (Figure 4b) and picture naming (Figure 5b) paradigms. This particular row of patients was selected for visual comparison of the other paradigms because they exemplified a variety of structural and functional patterns.

With the a priori analysis parameter set, the mean Dice coefficient of similarity across patients for the semantic paradigm was 0.66 ± 0.15 (range 0.40–0.82; Figure 6a), which was higher than the Dice coefficients for the narrative (0.47 ± 0.17; $|t(15)| = 4.93$; $p = .0002$; Cohen's $d_z = 1.23$) and picture naming (0.43 ± 0.17; $|t(15)| = 4.98$; $p = .0002$; Cohen's $d_z = 1.24$) paradigms, indicating that the semantic paradigm yielded the most reproducible activation maps.

To explore the generality of this finding, the mean Dice coefficient was then plotted for each paradigm as a function of ROI, voxelwise threshold, and cluster extent threshold (Figure 7a). This analysis showed that the semantic paradigm yielded Dice coefficients in the good range under many different parameter combinations. Many of these parameter sets also demonstrated strong evidence for validity (indicated with white rectangles), as will be described below. In contrast, Dice coefficients for the other two paradigms were mostly in the fair range. For these paradigms, activation maps tended to be more reproducible when the ROI was more circumscribed and when voxelwise thresholds were more lenient. However, these parameter sets did not yield good evidence for validity, as will be described below.

3.4 | Lateralization of language maps

A valid language mapping paradigm is expected to yield left-lateralized activation maps in the majority of neurologically normal participants. Activation maps for the adaptive semantic matching paradigm in each of the 14 neurologically normal participants are shown in Figure 8. Activations were clearly lateralized to the left hemisphere in 12 of 14 participants, were somewhat left-lateralized in one participant (NN13), and surprisingly, were right-lateralized in another participant (NN14).

The distribution of LIs on the three paradigms in the neurologically normal group is shown in Figure 6b. The individual with right-lateralized activations on the semantic task (NN14) is indicated with black dots. Importantly, NN14 also showed the most rightward lateralization of activation on the narrative comprehension paradigm. Accordingly, we assumed that she actually did have right hemisphere dominance for language, and so for the purpose of all subsequent analyses, her images were mirror-reversed. This means that in NN14, rightward lateralization was interpreted as reflecting correct lateralization (evidence for validity), and when evaluating sensitivity, activations in right rather than left frontal and temporal regions were quantified.

With the a priori analysis parameter set, the mean LI in controls for the semantic paradigm was 0.81 ± 0.24 (range 0.26–1.00), which was higher than the LIs for the narrative (0.34 ± 0.38; $|t(13)| = 6.83$; $p < .0001$; Cohen's $d_z = 1.82$) and picture naming (0.11 ± 0.20;
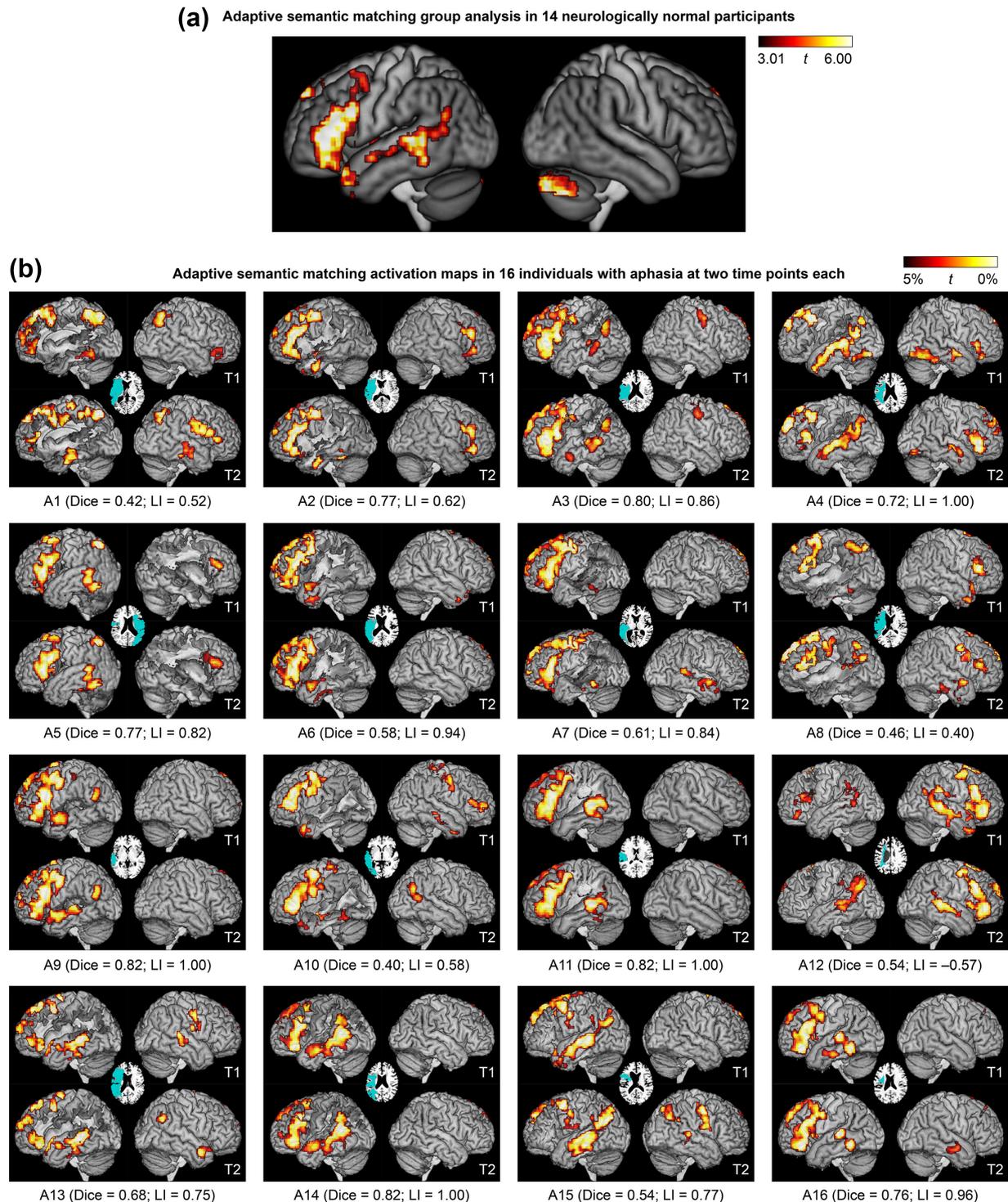
(a) Adaptive semantic matching group analysis in 14 neurologically normal participants

3.01 *t* 6.00

(b) Adaptive semantic matching activation maps in 16 individuals with aphasia at two time points each

5% *t* 0%

A1 (Dice = 0.42; LI = 0.52) A2 (Dice = 0.77; LI = 0.62) A3 (Dice = 0.80; LI = 0.86) A4 (Dice = 0.72; LI = 1.00)

A5 (Dice = 0.77; LI = 0.82) A6 (Dice = 0.58; LI = 0.94) A7 (Dice = 0.61; LI = 0.84) A8 (Dice = 0.46; LI = 0.40)

A9 (Dice = 0.82; LI = 1.00) A10 (Dice = 0.40; LI = 0.58) A11 (Dice = 0.82; LI = 1.00) A12 (Dice = 0.54; LI = −0.57)

A13 (Dice = 0.68; LI = 0.75) A14 (Dice = 0.82; LI = 1.00) A15 (Dice = 0.54; LI = 0.77) A16 (Dice = 0.76; LI = 0.96)

FIGURE 3 Language activation maps derived from the adaptive semantic matching paradigm. (a) Group analysis in 14 neurologically normal participants. Whole brain activations were thresholded at voxelwise $p < .005$ then corrected for multiple comparisons at $p < .01$ based on cluster extent. (b) Activation maps in 16 individuals with aphasia at two time points each. The patients are arranged in groups according to clinical impression, then in ascending order of overall QAB score within each group. See Participants section, and for more detailed language data, see Figure 14 of Wilson et al. (2018), which is laid out the same way. Voxels with the highest 5% of *t* statistics were plotted, subject to a minimum cluster volume of 2 cm^3, in an ROI comprising known language regions or plausible candidate regions for functional reorganization (*Lang+* ROI); note that the cerebellum was not included (unlike panel a). Inset axial slices show lesion reconstructions. T1 = first imaging session; T2 = second imaging session; Dice = Dice coefficient of similarity; LI = lateralization index

TABLE 4 Brain regions activated by the adaptive semantic matching paradigm in neurologically normal participants

Brain region(s)	MNI coordinates			Extent (mm³)	Max t	p
	x	y	z			
Left inferior frontal gyrus (pars opercularis, triangularis and orbitalis); precentral sulcus; posterior middle frontal gyrus; temporal pole	−45.4	25.5	8.5	31,784	10.28	<.001
Left superior temporal sulcus; middle temporal gyrus; angular gyrus; inferior temporal gyrus; fusiform gyrus; hippocampus	−42.6	−28.2	−6.4	27,000	8.53	<.001
Right cerebellum	25.3	−76	−32.8	13,104	11.48	<.001
Left superior frontal gyrus	−6.7	47.7	42.6	1,976	7.40	.002
Right hippocampus	24.3	−9.5	−16.7	1,904	7.01	.002

MNI coordinates show centers of mass.

$|t(13)| = 10.39$; $p < .0001$; Cohen's $d_z = 2.78$) paradigms. This indicates that the semantic paradigm yielded the most lateralized activation maps.

In an exploratory analysis, the distribution of LIs across the three paradigms was also investigated in individuals with aphasia (Figure 6b). This analysis was exploratory because there are no clear predictions as to what valid language maps should look like in individuals with aphasia. This analysis showed that activation maps derived from the semantic paradigm were clearly left-lateralized in 15 out of 16 patients (see also Figure 3b). The only exception was participant A12, whose LIs are indicated with black dots in Figure 6b. This patient also had the second-most right-lateralized activation map for the narrative paradigm, and the most right-lateralized map for the picture naming paradigm. Right lateralization for all three paradigms was reproduced across both sessions (Figures 3b, 4b, and 5b). Accordingly, we assumed that this patient actually did have right hemisphere dominance for language (whether or not due to reorganization), and so for the purpose of subsequent analyses, his images were flipped, just as for NN14.

The mean LI in patients for the semantic paradigm was 0.81 ± 0.22 (range 0.38–1.00), which was higher than the LIs for the narrative (0.26 ± 0.43; $|t(15)| = 7.59$; $p < .0001$; Cohen's $d_z = 1.89$) and picture naming (0.00 ± 0.34; $|t(15)| = 10.28$; $p < .0001$; Cohen's $d_z = 2.57$) paradigms, indicating that the semantic paradigm yielded the most lateralized activation maps in individuals with aphasia, just as it did in the neurologically normal group. We also investigated the test-retest reproducibility of LI in patients: intraclass correlation coefficients (type A-1) were excellent for the semantic ($r = .88$) and narrative ($r = .83$) paradigms, but poor for the picture naming paradigm ($r = .38$).

To explore the generality of these findings, mean LIs were then plotted for each paradigm as a function of ROI, voxelwise threshold, and cluster extent threshold in neurologically normal participants (Figure 7b) and individuals with aphasia (Figure 7c). This analysis showed that the semantic paradigm yielded lateralized activation maps under many different parameter combinations, many of which also showed good reliability and sensitivity (white rectangles). Remarkably similar patterns were seen in the patient group. The narrative paradigm yielded lateralized activation maps under many sets of parameters, albeit always to a lesser extent than the semantic paradigm. The

highest LIs arose when voxelwise thresholds were relative and stringent; these parameter sets did not, however, have good reliability. Individuals with aphasia showed less strongly lateralized maps than neurologically normal participants, which follows from the fact that activations are more bilateral for this paradigm, so when the left hemisphere was damaged, the left hemisphere component of the activation map was decreased, while the right hemisphere component remained intact. The picture naming paradigm generally resulted in largely symmetrical language maps, except when the ROI comprised only known language regions and the voxelwise threshold was stringent. Test-retest reproducibility under those circumstances was poor.
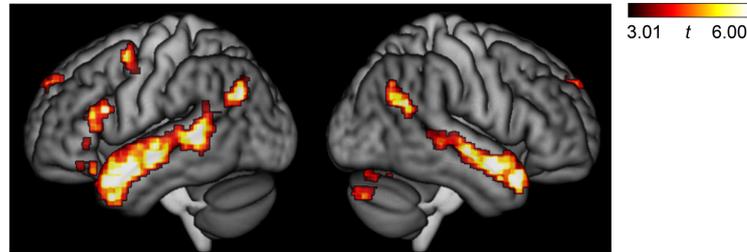
3.5 | Detection of frontal and temporal language areas

Besides revealing lateralized activation maps, a valid language mapping paradigm should activate dominant hemisphere frontal and temporal regions in essentially all neurologically normal participants. Therefore we compared the extent of activations in these regions across the three paradigms.

In neurologically normal participants, the extent of dominant hemisphere inferior frontal activation for the semantic paradigm was 16.0 ± 4.0 cm³ (range 10.5–22.4 cm³), which was greater than the activation extent for the narrative (4.6 ± 3.8 cm³; $|t(13)| = 8.21$; $p < .0001$; Cohen's $d_z = 2.19$) or picture naming (3.4 ± 2.5 cm³; $|t(13)| = 9.72$; $p < .0001$; Cohen's $d_z = 2.60$) paradigms (Figure 6c). The semantic paradigm yielded substantial frontal activations in all participants (Figure 8), indicating that the dominant frontal region was detected without fail in this cohort, while the narrative and picture naming paradigms yielded small or no activations in some participants (Figure 6c), suggesting a lack of sensitivity.

The extent of dominant hemisphere temporal activation for the semantic paradigm was 13.8 ± 3.6 cm³ (range 8.4–19.1 cm³), which did not differ from the extent for narrative (16.1 ± 5.1 cm³; $|t(13)| = 1.69$; $p = .11$; Cohen's $d_z = -0.45$), but was greater than the extent for naming (5.5 ± 3.9 cm³; $|t(13)| = 5.80$; $p < .0001$; Cohen's $d_z = 1.55$; Figure 6d). The semantic and narrative paradigms produced substantial temporal activations in all participants (semantic: Figure 8), indicating that the dominant temporal region was detected without fail in this cohort, while the picture naming paradigm did not.

(a) **Narrative comprehension group analysis in 14 neurologically normal participants**

(b) **Narrative comprehension activation maps in 4 individuals with aphasia at two time points each**

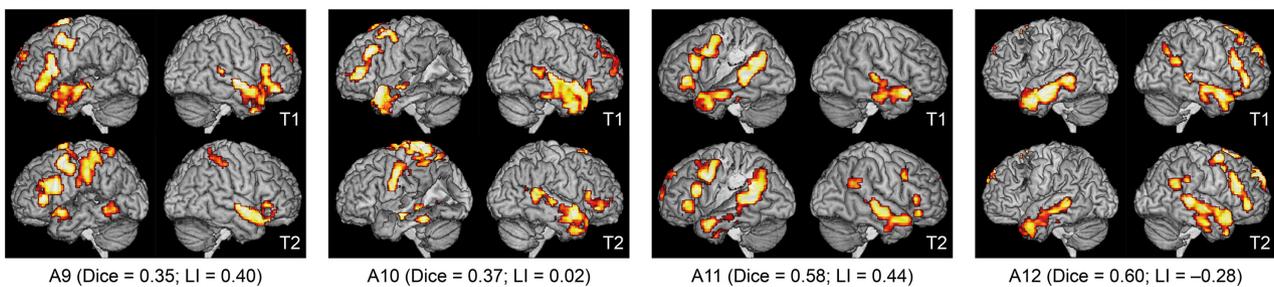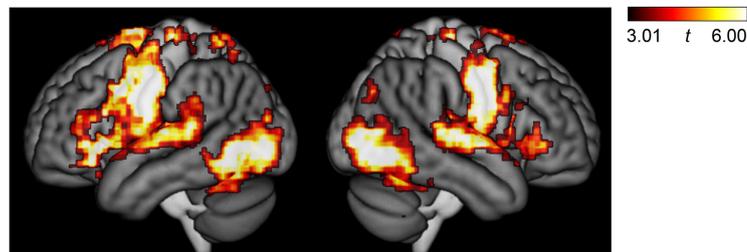A9 (Dice = 0.35; LI = 0.40)   A10 (Dice = 0.37; LI = 0.02)   A11 (Dice = 0.58; LI = 0.44)   A12 (Dice = 0.60; LI = −0.28)

FIGURE 4 Language activation maps derived from the narrative comprehension paradigm. (a) Group analysis in 14 neurologically normal participants. (b) Activation maps in 4 of the 16 individuals with aphasia at two time points each. These 4 patients are the 4 patients in the third row of Figure 3. See Figure 3 caption for additional definitions and explanations

In individuals with aphasia, there was no expectation that either language region should necessarily be present, due to structural damage as well as possible functional changes. We found that the semantic paradigm yielded frontal activation with extent within or slightly above the range of activation extents of the neurologically normal participants in 12 out of 16 patients (including right-lateralized A12; Figure 6c); in 10 of these patients, the inferior frontal gyrus was intact and in 2 it

was partially damaged (A2 and A14; Figure 3b). Four patients (A1, A8, A13, and A15) showed reduced frontal activation that fell outside the normal range (Figure 6c). The inferior frontal gyrus was largely destroyed in two of these cases (A1 and A13) and substantially damaged in the other two (A8 and A15; Figure 3b). There was perilesional activation in all four cases (Figure 3b). The narrative and naming paradigms yielded similar extents of frontal activation in individuals with

(a) **Picture naming group analysis in 14 neurologically normal participants**

(b) **Picture naming activation maps in 4 individuals with aphasia at two time points each**

A9 (Dice = 0.71; LI = 0.18)   A10 (Dice = 0.25; LI = −0.13)   A11 (Dice = 0.45; LI = −0.20)   A12 (Dice = 0.39; LI = −0.65)
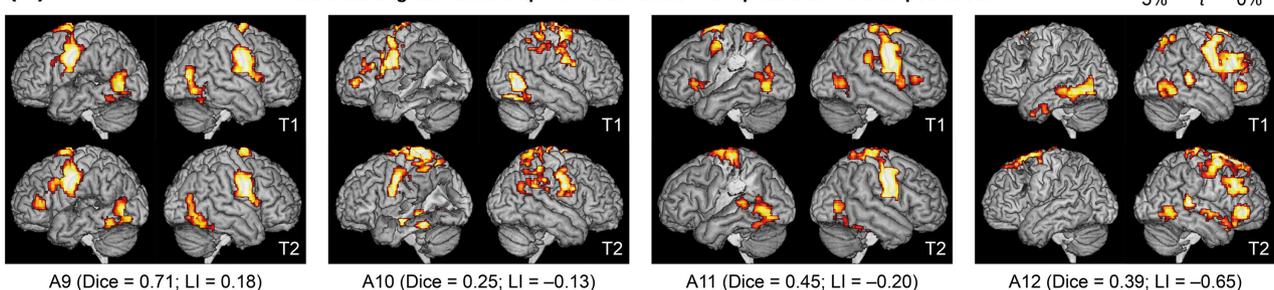
FIGURE 5 Language activation maps derived from the picture naming paradigm. (a) Group analysis in 14 neurologically normal participants. (b) Activation maps in 4 of the 16 individuals with aphasia at two time points each. These 4 patients are the 4 patients in the third row of Figure 3. See Figure 3 caption for additional definitions and explanations

aphasia and neurologically normal participants, suggesting a lack of sensitivity to detect reduced frontal activation (Figure 6c).

The semantic paradigm yielded temporal activation with extent within or slightly above the normal range in 8 patients (A3, A4, A9, A11, A12 [right-lateralized], A13, A14, and A15) and close to the normal range and typically localized in 2 patients (A5 and A16; Figure 6d).
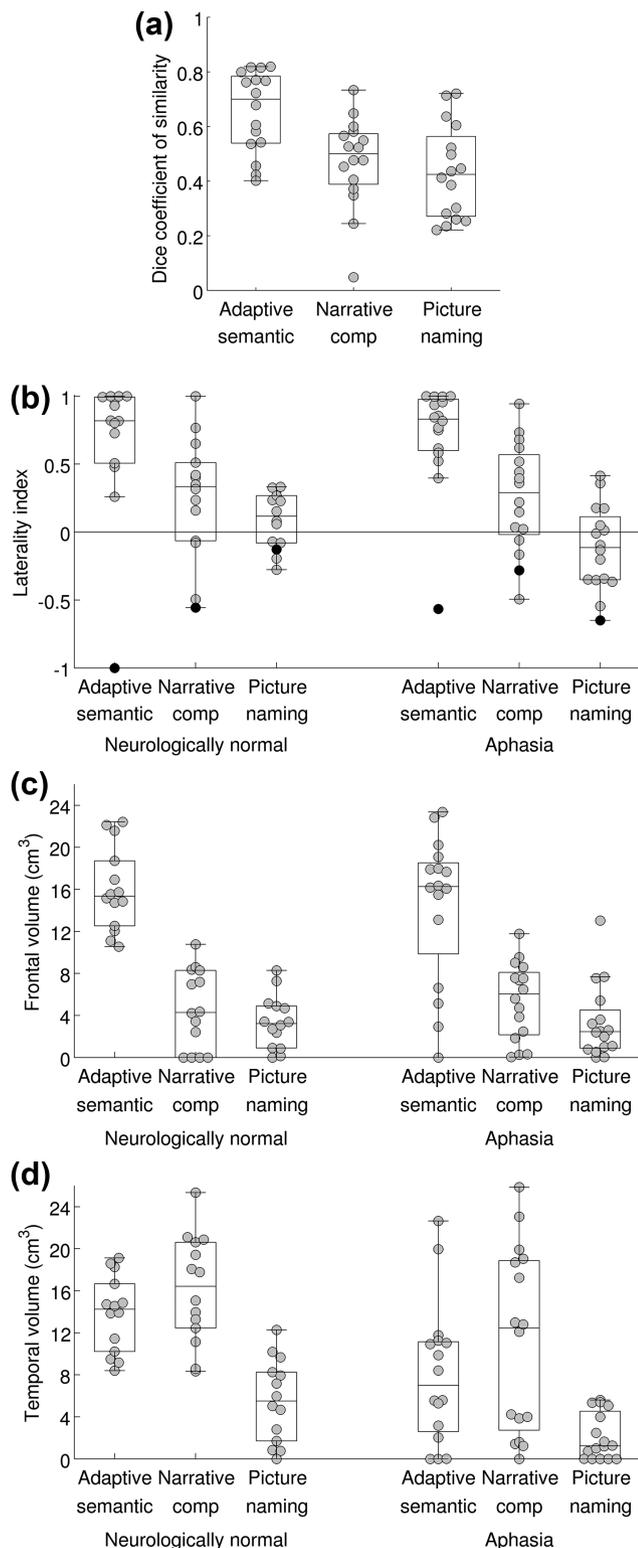

(a)

(b)

(c)

(d)

FIGURE 6.

In these 10 patients, the temporal language ROI was intact in 7 patients and partially damaged in 3 (A9, A13, and A14; Figure 3b). There were 5 patients with markedly reduced temporal activation (A2, A6, A7, A8, and A10) and 1 with temporal activation close to the normal extent but atypically localized (A1; Figure 6d). The temporal language region was severely damaged in all of these six patients (Figure 3b). All showed perilesional activation (Figure 3b). The narrative paradigm showed 9 patients within or slightly above the normal range and 7 below the normal range (A1, A2, A6, A7, A8, A9, and A10; Figure 6d), including all 6 who showed abnormal activation with the semantic paradigm. This supports the validity of both paradigms with respect to detection of reduced or abnormal temporal activation. The picture naming paradigm showed somewhat reduced temporal activation in patients, but temporal activations were not consistent in neurologically normal participants so this could not be interpreted (Figure 6d).

The impact of ROI, voxelwise threshold, and cluster volume cutoff on sensitivity to detect frontal and temporal language regions is described in Supporting Information online (Figures S1 and S2).

3.6 | Revised adaptive semantic stimuli

A second group of neurologically normal individuals was scanned on the revised version of the adaptive semantic matching paradigm, which was modified to minimize ceiling effects in people without aphasia. This second group still performed slightly better on the revised semantic task (87.3% ± 4.0%) than on the revised perceptual task (85.2% ± 2.6% accuracy; $|t(15)| = 2.20$; $p = .044$), but the ceiling effect on the semantic task was greatly ameliorated (Figure 9a). The increased difficulty of the revised paradigm at harder levels should not pose problems for individuals with aphasia, since the difficulty of the easier levels was not increased.

The mean difficulty level of items presented was greater for semantic trials (5.09 + 0.67) than perceptual trials (4.56 + 0.52; $|t(15)| = 3.26$; $p = .0052$; Figure 9a), but this is not especially important since the two tasks are not inherently matched across specific difficulty

FIGURE 6 Psychometric assessment of the three paradigms. (a) Test–retest reproducibility of the three paradigms in individuals with aphasia. The distribution of the Dice coefficient is plotted (relative voxelwise threshold = 5%; minimum cluster volume = 2 cm^3; ROI = Lang+, that is language regions and plausible candidates for reorganization). (b) Lateralization of language maps. The distribution of the lateralization index is plotted for neurologically normal individuals, and for individuals with aphasia, for each of the three paradigms. Individuals with apparent right hemisphere language dominance are indicated with black dots (NN14 and A12). (c) Sensitivity for detection of dominant hemisphere frontal language area. The distribution of the activation volume in the dominant inferior frontal gyrus is plotted for neurologically normal individuals, and for individuals with aphasia, for each of the three paradigms. (d) Sensitivity for detection of dominant hemisphere temporal language area. The distribution of the activation volume in the temporal ROI is plotted for neurologically normal individuals, and for individuals with aphasia, for each of the three paradigms
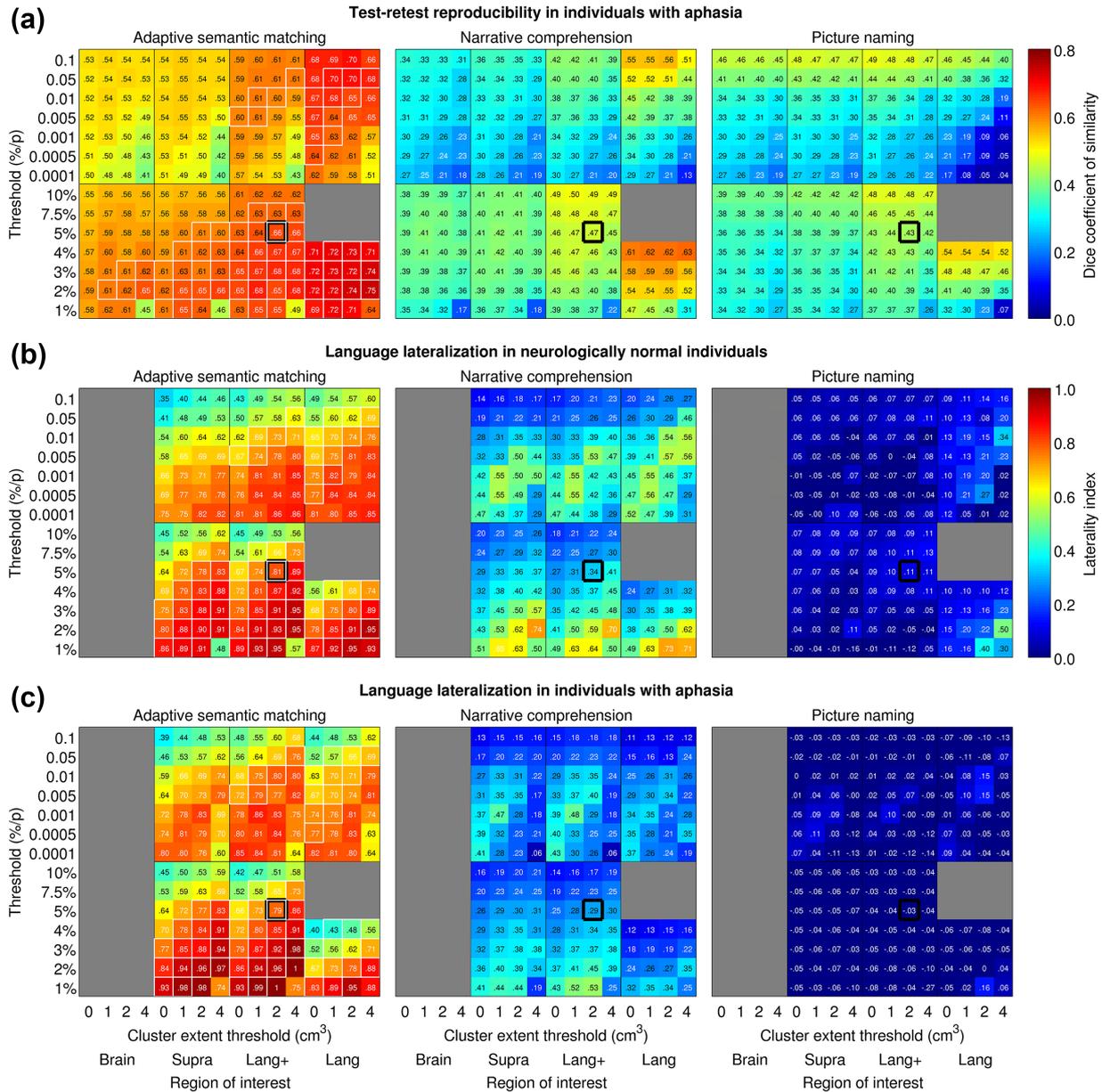
FIGURE 7 Impact of analysis parameter sets. (a) Impact of analysis parameters on test–retest reproducibility of the three paradigms in individuals with aphasia. Mean Dice coefficients of similarity across participants are plotted as a function of absolute and relative voxelwise thresholds (*y* axes), region of interest (*x* axes) and minimum cluster volume (*x* axes). Regions of interest: *Brain* = whole brain; *Supra* = supratentorial cortical regions; *Lang*+ = known language regions and plausible candidates for reorganization; *Lang* = known language regions and their homotopic counterparts. White outlines indicate parameter sets that exemplified desirable psychometric properties across the board (Dice ≥ 0.60; LI ≥ 0.60; frontal and temporal regions detected without fail in the neurologically normal group). Thick black outlines show the *a priori* analysis parameter set. (b) Impact of analysis parameters on language lateralization in neurologically normal participants. Mean lateralization indices across participants are plotted for each paradigm as a function of absolute and relative voxelwise thresholds (*y* axes), region of interest (*x* axes) and minimum cluster volume (*x* axes). (c) Impact of analysis parameters on language lateralization in individuals with aphasia

levels. Reaction times on correct trials were faster on semantic trials (1,465 + 172 ms) than perceptual trials (1,861 + 159 ms; |*t* (15)| = 7.82; *p* < .0001; Figure 9a). While it would be preferable for reaction times to be equivalent, this is not a serious concern as explained earlier.

In this neurologically normal group, the contrast between the semantic and perceptual conditions activated a similar set of brain

regions to the original version of the paradigm, including the left inferior frontal gyrus (pars opercularis, triangularis and orbitalis) and posterior superior temporal gyrus, superior temporal sulcus and middle temporal gyrus (Figure 9b). Compared to the original version of the paradigm, there appeared to be more robust activation of left anterior temporal cortex and right inferior frontal cortex, which might be expected given the increased difficulty of the task, but the data from
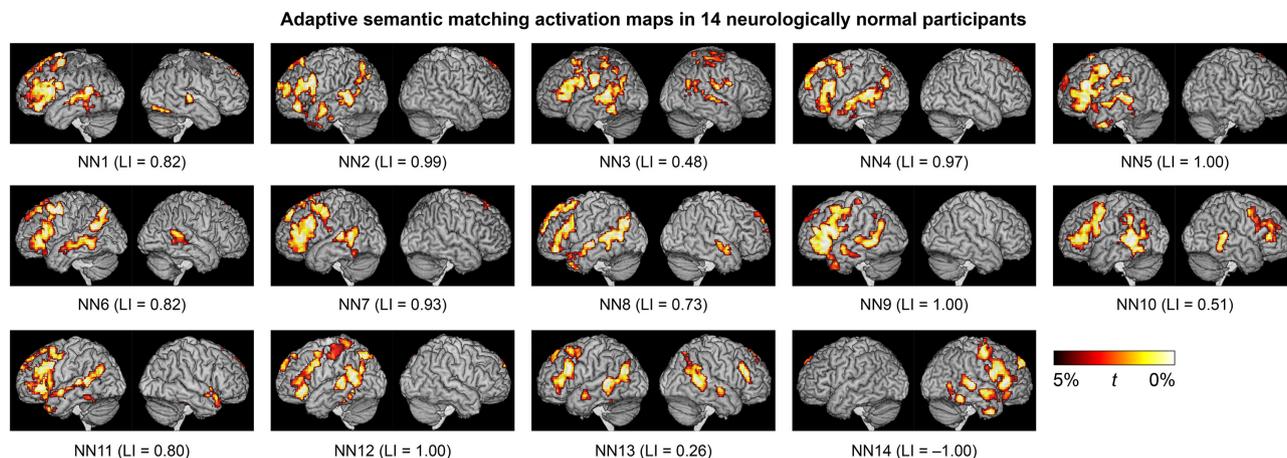
Adaptive semantic matching activation maps in 14 neurologically normal participants

NN1 (LI = 0.82) NN2 (LI = 0.99) NN3 (LI = 0.48) NN4 (LI = 0.97) NN5 (LI = 1.00)

NN6 (LI = 0.82) NN7 (LI = 0.93) NN8 (LI = 0.73) NN9 (LI = 1.00) NN10 (LI = 0.51)

NN11 (LI = 0.80) NN12 (LI = 1.00) NN13 (LI = 0.26) NN14 (LI = −1.00)

5% *t* 0%

FIGURE 8 Language activation maps derived from the adaptive semantic matching paradigm in 14 neurologically normal individuals. See Figure 3 caption for more information

4 | DISCUSSION

Our findings show that the adaptive semantic matching paradigm is appropriate for individuals with aphasia, has good test-retest reproducibility, and consistently identifies lateralized frontal and temporal language regions. The suitability of the paradigm for people with aphasia was demonstrated by the ability of all 16 participants to learn the task and perform above chance in the scanner. Test-retest reproducibility was quantified in terms of the Dice coefficient of similarity, which was

higher for the adaptive semantic matching paradigm than the two comparison paradigms, and was good over a wide range of potential analysis parameter sets. Validity was demonstrated by showing that the adaptive semantic matching paradigm yields more lateralized language maps than the two comparison paradigms, and identifies frontal and temporal language regions with high sensitivity over a wide range of analysis parameter sets.

4.1 | Design features underlying the feasibility, reliability and validity of the semantic paradigm

The most important feature that makes the semantic paradigm suitable for individuals with aphasia is its adaptive nature, which permits the same paradigm to be used to map language regions in people with different degrees of language impairment, as well as individuals with normal language function. The difficulty of language items is manipulated in terms of lexical frequency, concreteness, word length, age of acquisition of words, degree of relatedness of word pairs, and presentation rate, all of which are soundly established factors that impact the difficulty of linguistic tasks (Coltheart, 1981). The difficulty of the perceptual control task is also adaptively manipulated, ensuring that both tasks remained similarly challenging. Some previous developmental studies have used language mapping paradigms with different predetermined levels of difficulty for children of different age ranges (Berl et al., 2010, 2014; Gaillard et al., 2007; You et al., 2011), and in some studies of aphasia recovery, presentation rates have been adapted for individual patients (e.g. Heiss et al., 1999). Our adaptive semantic paradigm goes beyond these previous approaches in three ways: first, by dynamically selecting stimuli that are uniquely adapted to each individual, second, by manipulating multiple theoretically motivated linguistic variables, and third, by also adjusting the difficulty of the control condition, which was not varied in previous studies.

Another simple but important feature is the use of just a single button by which participants identify "match" trials, such that mismatch trials require no response. This avoids participants having to learn an arbitrary association between responses and buttons, which can prove
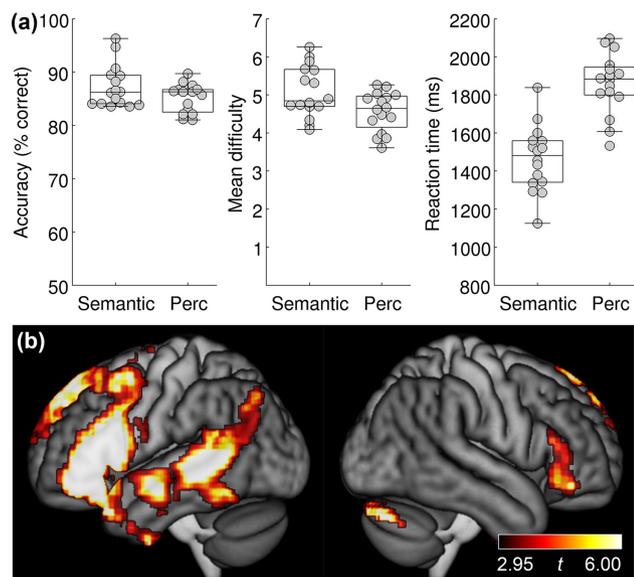

FIGURE 9 Revised adaptive semantic matching paradigm. (a) Accuracy, item difficulty, and reaction time on the revised semantic and perceptual control conditions in a second group of 16 neurologically normal participants. Perc = Perceptual. (b) Group analysis in this neurologically normal group. Whole brain activations were thresholded at voxelwise $p < .005$ then corrected for multiple comparisons at $p < .01$ based on cluster extent

difficult for some neurological patients. Furthermore, the multiple phases of training—a flexible examiner-guided phase, an independent phase, and another independent phase in the scanner before acquisition of functional data begins—are important in ensuring that all participants can perform the task.

The test-retest reproducibility of the adaptive semantic matching paradigm was good, with a mean Dice coefficient of similarity of 0.66 with our *a priori* analysis parameter set, and even higher Dice coefficients with other parameter sets. This compares favorably to all other reliability coefficients that have been reported for language mapping paradigms (Billingsley-Marshall, Simos, & Papanicolaou, 2004; Brannen et al., 2001; Fernández et al., 2003; Fesl et al., 2010; Gross & Binder, 2014; Harrington et al., 2006; Maldjian et al., 2002; Rau et al., 2007; Rutten et al., 2002; see Wilson et al., 2017 for review); to our knowledge, the highest Dice coefficient previously reported for a paradigm with reasonable validity is 0.61 (Fesl et al., 2010), and participants in that study were neurologically normal. Other investigations of language mapping reliability in aphasia have used metrics such as the voxelwise intraclass correlation coefficients that do not provide an overall assessment of reliability (Eaton et al., 2008; Meltzer et al., 2009).

Probably the most important factor underlying the good test–retest reproducibility of the semantic paradigm is the fact that both the semantic task and the perceptual task are highly constrained in terms of the linguistic and other processing required. This ensures that similar cognitive states in both conditions are induced each time participants perform the task. In comparison, a contrast such as narrative comprehension versus backwards speech may lead to very different results on different occasions depending on how attention is modulated by the extent to which the participant is interested in the narrative, the extent to which they are interested in the backwards speech, and so on (Bautista & Wilson, 2016; Wild, Davis, & Johnsrude, 2012). Many other factors are likely to have contributed to the reliability observed, such as a preprocessing pipeline optimized for neurological patients (e.g., manual removal of noise components); judicious use of individual gray matter and lesion masks; and systematic consideration of a range of plausible regions of interest, absolute and relative voxelwise thresholds, and cluster volume cutoffs, which can have a large impact on reliability (Bennett & Miller, 2010; Wilson et al., 2017).

Validity was assessed in terms of each paradigm's ability to reveal known features of normal language organization: lateralization, and activation of frontal and temporal regions (Knecht et al., 2003; Seghier et al., 2011; Springer et al., 1999; Tzourio-Mazoyer et al., 2010; Bradshaw et al., 2017). The strongly lateralized activation maps derived from the semantic paradigm are likely driven by the fact that it is an active, rather than a passive, task (Vannest et al., 2009). That is, it requires specific linguistic processing and overt responses. In contrast, passive tasks such as narrative comprehension yield less lateralized activation patterns (Maldjian et al., 2002; Crinion et al., 2003; Harrington et al., 2006), in accordance with the fact that language comprehension is more bilaterally represented than language production (Hickok & Poeppel, 2007). This is not to imply that the less lateralized language maps derived from the narrative comprehension paradigm are "wrong" per se. But for the purpose of investigating language reorganization

after damage to language regions, a paradigm that highlights lateralized aspects of the language network seems to offer the most potential for documenting any functional reorganization, because aphasia so consistently follows from damage to the left, but not the right hemisphere. Note that picture naming is also an active task, but presumably the active and lateralized function of lexical access is so effortless in most cases that activation related to it is swamped by bilateral activations to sensorimotor aspects of the task.

The high sensitivity of the adaptive semantic matching paradigm to both frontal as well as temporal language regions is likely due to the fact that it implicates both active semantic decision making and comprehension processes (Mbwana et al., 2009). A similar but non-adaptive contrast in a well-powered previous study activated the temporal language region much less robustly than we observed (Seghier et al., 2004), suggesting that the challenging processing elicited by the adaptive nature of our paradigm may contribute to its sensitivity. While the narrative paradigm also had excellent temporal sensitivity, it did not have good frontal sensitivity, and neither region was identified consistently by the picture naming paradigm; these findings are consistent with prior research (Rau et al., 2007; Wilson et al., 2017).

4.2 | Limitations

Despite the overall success of the adaptive semantic matching paradigm, several important limitations should be noted. First, the semantic paradigm identifies only one component of the language network: specifically, the lateralized frontal and temporal regions. While these two lateralized regions arguably constitute the core of the language network, there are several other interacting networks that are also critical for language function. As mentioned earlier, language comprehension is supported not only by these lateralized regions, but also by bilateral regions in the anterior temporal lobe and angular gyrus (Binder, Desai, Graves, & Conant, 2009; Crinion & Price, 2005; Hickok & Poeppel, 2007; Patterson, Nestor, & Rogers, 2007; Rice, Hoffman, & Lambon Ralph, 2015). In addition, there are other lateralized fronto-parietal networks supporting other language processes such as phonological retrieval and encoding (Pillay, Stengel, Humphries, Book, & Binder, 2014) and motor speech programming (Graff-Radford et al., 2014). In a future publication, we will describe two adaptive phonological tasks that attempt to identify some of these other regions.

Second, despite our best efforts to design a maximally simple paradigm, not all individuals with aphasia will be able to perform the adaptive semantic matching task. In our cohort of 16 people with chronic post-stroke aphasia, drawn mostly from a community aphasia group, all but one patient had excellent comprehension of written single words, and all patients demonstrated preserved non-verbal semantic processing. If either or both of these functions were impaired, as may be expected in syndromes such as global aphasia, Wernicke's aphasia, or semantic dementia, then task performance would clearly be impacted. Indeed, the one patient with impaired single word comprehension showed a floor effect on the semantic task (though his performance was still better than chance).

Third, and related to this observation, the adaptive design was not entirely successful in equating performance across the semantic and perceptual tasks. Not only was there a floor effect in the semantic condition in one patient, but in the original version of the paradigm (on which most of the data were acquired), there were also ceiling effects for all but one of the neurologically normal participants, and even one of the individuals with aphasia (notably, the one who had almost completely recovered). The revised paradigm, intended for future use, largely resolved this problem. However floor effects will likely continue to pose problems in individuals with more impaired language; if these individuals are not performing at the same accuracy rate as others, then this would be a limitation when comparing their language maps to less impaired patients or neurologically normal individuals. Finally it should be noted that the yoking of presentation rate across the two conditions, which was necessary to counterbalance sensorimotor aspects of the tasks, will inherently interfere with the independence of the interweaved adaptive staircase procedures.

Fourth, while a mean Dice coefficient of similarity of 0.66 reflects very satisfactory test–retest reproducibility as can be readily appreciated in Figure 3, there were nevertheless sometimes still significant discrepancies between the language maps obtained in the two sessions. For example, patient A1 had a much more prominent right frontal activation in the second session compared to the first session (this participant's Dice coefficient was 0.42, which was one of the lowest). Therefore, it is still necessary to be cautious in attributing observed changes in language activation maps to neuroplasticity. Approaches such as multiple sessions at each time point, and random effects group analyses where feasible, will still be important in documenting any functional reorganization that may take place (Kiran et al., 2013; Meinzer et al., 2013). It is also noteworthy that better reliability was obtained with relative rather than absolute thresholds. This makes sense, given that relative thresholds essentially factor out general activation strength, which is one important source of undesirable variability (Gross & Binder, 2014; Knecht et al., 2003; Voyvodic, 2012; Wilson et al, 2017). However, the downside of relative thresholds is that any genuine changes in the extent of cortical regions devoted to language processing would not be captured, which could be a significant limitation depending on one's hypotheses.

4.3 | Language localization in chronic post-stroke aphasia

While it was not the primary goal of this paper to address empirical questions regarding language localization in chronic aphasia, our findings were clear enough in some respects that some preliminary observations can be made.

First, the dominant hemisphere frontal and temporal regions that were activated in every neurologically normal individual were also activated in many individuals with aphasia. Specifically, twelve patients showed essentially normal frontal activations and ten showed normal or near-normal temporal activations (albeit reduced in extent in some cases). Generally these normally activated regions were structurally spared, but there were several cases in which partially damaged regions showed normal or near-normal activation, suggesting that partially damaged regions may often retain their functional role in language processing. This basic preservation of language activation patterns in chronic post-stroke aphasia is consistent with prior research (Griffis et al., 2017; Heiss & Thiel, 2006; Saur et al., 2006).

Second, four patients showed reduced or absent inferior frontal activations, and six patients showed reduced, absent or aberrant temporal activations. In all of these cases, the structures in question were substantially damaged or destroyed, and in all cases perilesional activation was observed. It is not yet clear whether these perilesional activations represent the outcome of functional reorganization (Robson et al., 2014), or whether these perilesional regions would have been involved in language processing premorbidly, since there was considerable variability in neurologically normal participants in terms of the extent and precise localization of frontal and temporal activations.

Third, it is apparent that reorganization of language processing to the right hemisphere (Weiller et al., 1995; Turkeltaub et al., 2011) is an uncommon outcome at best. Only one individual with aphasia (A12) showed right-lateralized language regions. It cannot be determined whether this represents the outcome of reorganization, or whether he had premorbid right-lateralized or bilateral language function. Given that his left hemisphere stroke did result in aphasia, right hemisphere dominance for language seems unlikely. But bilateral language, similar to neurologically normal participant NN13, is a distinct possibility. The absence of any strong evidence for reorganization to the right hemisphere in any other patient is especially striking in view of how well recovered many language functions were relative to lesion location. In particular, four of the five individuals with conduction aphasia (A6, A7, A8, and A10) had large left posterior temporal lesions and none had normal temporal functional activation. Yet all of these participants had excellent single word comprehension and semantic processing abilities, as evidenced by their task performance, as well as language and neuropsychological data (Wilson et al., 2018). They also were fluent, could produce sentences (albeit with paragrammatic and paraphasic errors), and were able to retrieve words, repeat words and sentences, and read aloud, to varying extents (Wilson et al., 2018). All of these functions normally depend on the posterior left temporal language area, but show quite rapid recovery after this region is damaged (Kertesz, Lau, & Polk, 1993; Naeser, Helm-Estabrooks, Haas, Auerbach, & Srinivasan, 1987; Selnes, Knopman, Niccum, Rubens, & Larson, 1983; Selnes, Niccum, Knopman, & Rubens, 1984; Yagata et al., 2017). There was little evidence for comparable activation of the homotopic right hemisphere region, even though it would be a plausible substrate for reorganization given the relative bilaterality of temporal lobe language regions to begin with (Hickok & Poeppel, 2007). While these four patients each showed small right temporal lobe activations in one or both sessions, these activations were limited in extent, were mostly not homotopic to the left posterior temporal language region, and their specific locations were never replicated across the two sessions.

Why then is language function as good as it is in these individuals? There are several possibilities, which cannot yet be distinguished between. It may be that small and displaced residual left hemisphere temporal and/or parietal activations that each of these participants

evidenced (e.g., anterior temporal in A6, ventral temporal in A7, angular gyrus in A8, etc.) are sufficient to support the relevant language functions to the extent that they are spared. Another possibility is that right temporal cortex is in fact now critical for language function, but is not activated by the semantic task for some unknown reason, though this would of course be difficult to explain given the robust activation of the left temporal cortex by the semantic contrast in neurologically normal individuals. Addressing these kinds of questions will require recruiting larger groups of individuals with aphasia so that relationships between functional activation of different regions, and patterns of spared and impaired language function, can be investigated.

5 | CONCLUSION

The adaptive semantic matching paradigm described in this paper is appropriate for individuals with aphasia, has good test-retest reproducibility, and successfully identifies lateralized frontal and temporal language regions. Accordingly, it should prove to be a useful tool for research on neuroplasticity in recovery from aphasia. Moreover, the paradigm also has a potential application in presurgical language mapping (Binder et al., 2008), especially in the minority of presurgical patients who present with significant language deficits. Given the substantial psychometric differences we observed between paradigms, it is clear that future language mapping paradigms should be carefully assessed in terms of feasibility, reliability and validity.

ORCID

Stephen M. Wilson http://orcid.org/0000-0001-9884-2852
Melodie Yen http://orcid.org/0000-0002-9257-1147

REFERENCES

Abel, S., Weiller, C., Huber, W., Willmes, K., & Specht, K. (2015). Therapy-induced brain reorganization patterns in aphasia. *Brain*, *138*(4), 1097–1112.

Allendorfer, J. B., Kissela, B. M., Holland, S. K., & Szaflarski, J. P. (2012). Different patterns of language activation in post-stroke aphasia are detected by overt and covert versions of the verb generation fMRI task. *Medical Science Monitor*, *18*(3), CR135–CR147.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–851.

Bautista, A., & Wilson, S. M. (2016). Neural responses to grammatically and lexically degraded speech. *Language, Cognition and Neuroscience*, *31*(4), 567–574.

Beckmann, C. F., & Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, *23*(2), 137–152.

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.

Berl, M. M., Duke, E. S., Mayo, J., Rosenberger, L. R., Moore, E. N., VanMeter, J., . . . Gaillard, W. D. (2010). Functional anatomy of listening and reading comprehension during development. *Brain and Language*, *114*(2), 115–125.

Berl, M. M., Zimmaro, L. A., Khan, O. I., Dustin, I., Ritzl, E., Duke, E. S., . . . Gaillard, W. D. (2014). Characterization of atypical language activation patterns in focal epilepsy. *Annals of Neurology*, *75*(1), 33–42.

Berthier, M. L., & Davila, G. (Eds.). (2014). *Pharmacology and aphasia*. New York: Routledge.

Billingsley-Marshall, R., Simos, P., & Papanicolaou, A. (2004). Reliability and validity of functional neuroimaging techniques for identifying language-critical areas in children and adults. *Developmental Neuropsychology*, *26*(2), 541–563.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796.

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*, *17*(1), 353–362.

Binder, J. R., Medler, D. A., Desai, R., Conant, L. L., & Liebenthal, E. (2005). Some neurophysiological constraints on models of word naming. *NeuroImage*, *27*(3), 677–693.

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*(12), 1980–1997.

Birn, R. M., Bandettini, P. A., Cox, R. W., & Shaker, R. (1999). Event-related fMRI of tasks involving brief motion. *Human Brain Mapping*, *7*(2), 106–114.

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

Brady, M. C., Kelly, H., Godwin, J., Enderby, P., & Campbell, P. (2016). Speech and language therapy for aphasia following stroke. *Cochrane Database of Systematic Reviews*, *6*, CD000425.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Brallier, J. (2002). *Who was Albert Einstein*? New York: Grosset & Dunlap.

Brannen, J. H., Badie, B., Moritz, C. H., Quigley, M., Meyerand, M. E., & Haughton, V. M. (2001). Reliability of functional MR imaging with word-generation tasks for mapping Broca's area. *American Journal of Neuroradiology*, *22*(9), 1711–1718.

Breining, B. L., Lala, T., Cuitiño, M. M., Manes, F., Peristeri, E., Tsapkini, K., . . . Hillis, A. E. (2015). A brief assessment of object semantics in primary progressive aphasia. *Aphasiology*, *29*(4), 488–505.

Brett, M., Leff, A. P., Rorden, C., & Ashburner, J. (2001). Spatial normalization of brain images with focal lesions using cost function masking. *NeuroImage*, *14*(2), 486–500.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *33* (4), 497–505.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173.

Crinion, J. T., Lambon Ralph, M. A., Warburton, E. A., Howard, D., & Wise, R. J. S. (2003). Temporal lobe regions engaged during normal speech comprehension. *Brain*, *126*(5), 1193–1201.

Crinion, J., & Price, C. J. (2005). Right anterior superior temporal activation predicts auditory sentence comprehension following aphasic stroke. *Brain*, *128*(12), 2858–2871.

Crinion, J. T., Warburton, E. A., Lambon Ralph, M. A., Howard, D., & Wise, R. J. S. (2006). Listening to narrative speech after aphasic stroke: The role of the left anterior temporal lobe. *Cerebral Cortex*, *16* (8), 1116–1125.

Eaton, K. P., Szaflarski, J. P., Altaye, M., Ball, A. L., Kissela, B. M., Banks, C., & Holland, S. K. (2008). Reliability of fMRI for studies of language in post-stroke aphasia subjects. *NeuroImage*, *41*(2), 311–322.

Edgers, G. (2006). *Who were the Beatles?* New York: Grosset & Dunlap.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences USA*, *113*(28), 7900–7905.

Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in Cognitive Sciences*, *18*(3), 120–126.

Fernández, G., de Greiff, A., von Oertzen, J., Reuber, M., Lun, S., Klaver, P., ... Elger, C. E. (2001). Language mapping in less than 15 minutes: Real-time functional MRI during routine clinical investigation. *NeuroImage*, *14*(3), 585–594.

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

Fesl, G., Bruhns, P., Rau, S., Wiesmann, M., Ilmberger, J., Kegel, G., & Brueckmann, H. (2010). Sensitivity and reliability of language laterality assessment with a free reversed association task—a fMRI study. *European Radiology*, *20*(3), 683–695.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state : A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198.

Fridriksson, J., Baker, J. M., & Moser, D. (2009). Cortical mapping of naming errors in aphasia. *Human Brain Mapping*, *30*(8), 2487–2498.

Fridriksson, J., Richardson, J. D., Fillmore, P., & Cai, B. (2012). Left hemisphere plasticity and aphasia recovery. *NeuroImage*, *60*(2), 854–863.

Gaillard, W. D., Berl, M. M., Moore, E. N., Ritzl, E. K., Rosenberger, L. R., Weinstein, S. L., ... Theodore, W. H. (2007). Atypical language in lesional and nonlesional complex partial epilepsy. *Neurology*, *69*(18), 1761–1771.

García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. *Vision Research*, *38* (12), 1861–1881.

Gelinas, J. N., Fitzpatrick, K. P. V., Kim, H. C., & Bjornson, B. H. (2014). Cerebellar language mapping and cerebral language dominance in pediatric epilepsy surgery patients. *NeuroImage: Clinical*, *6*, 296–306.

Geranmayeh, F., Brownsett, S. L. E., & Wise, R. J. S. (2014). Task-induced brain activity in aphasic stroke patients: What is driving recovery? *Brain*, *137*(10), 2632–2648.

Gitelman, D. R., Nobre, A. C., Sonty, S., Parrish, T. B., & Mesulam, M.-M. (2005). Language network specializations: An analysis with parallel task designs and functional magnetic resonance imaging. *NeuroImage*, *26*(4), 975–985.

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.

Graff-Radford, J., Jones, D. T., Strand, E. A., Rabinstein, A. A., Duffy, J. R., & Josephs, K. A. (2014). The neuroanatomy of pure apraxia of speech in stroke. *Brain and Language*, *129*, 43–46.

Griffis, J. C., Nenert, R., Allendorfer, J. B., Vannest, J., Holland, S., Dietz, A., & Szaflarski, J. P. (2017). The canonical semantic network supports residual language function in chronic post-stroke aphasia. *Human Brain Mapping*, *38*(3), 1636–1658.

Gross, W. L., & Binder, J. R. (2014). Alternative thresholding methods for fMRI data optimized for surgical planning. *NeuroImage*, *84*, 554–561.

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

Heiss, W. D., Kessler, J., Thiel, A., Ghaemi, M., & Karbe, H. (1999). Differential capacity of left and right hemispheric areas for compensation of poststroke aphasia. *Annals of Neurology*, *45*(4), 430–438.

Heiss, W.-D., & Thiel, A. (2006). A proposed regional hierarchy in recovery of post-stroke aphasia. *Brain and Language*, *98*(1), 118–123.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402.

Howard, D., & Patterson, K. (1992). *Pyramids and palm trees: A test of semantic access from pictures and words*. Bury St. Edmunds: Thames Valley.

Janecek, J. K., Swanson, S. J., Sabsevitz, D. S., Hammeke, T. A., Raghavan, M., E. Rozman, M., & Binder, J. R. (2013). Language lateralization by fMRI and Wada testing in 229 patients with epilepsy: Rates and predictors of discordance. *Epilepsia*, *54*(2), 314–322.

Jansen, A., Menke, R., Sommer, J., Förster, A. F., Bruchmann, S., Hempleman, J., ... Knecht, S. (2006). The assessment of hemispheric lateralization in functional MRI—Robustness and reproducibility. *NeuroImage*, *33*(1), 204–217.

Kelly, R. E., Alexopoulos, G. S., Wang, Z., Gunning, F. M., Murphy, C. F., Morimoto, S. S., ... Hoptman, M. J. (2010). Visual inspection of independent components: Defining a procedure for artifact removal from fMRI data. *Journal of Neuroscience Methods*, *189*(2), 233–245.

Kertesz, A. (1982). *Western Aphasia Battery*. New York: Grune and Stratton.

Kertesz, A., Lau, W. K., & Polk, M. (1993). The Structural determinants of recovery in Wernicke's aphasia. *Brain and Language*, *44*(2), 153–164.

Kertesz, A., & McCabe, P. (1977). Recovery patterns and prognosis in aphasia. *Brain*, *100*(1), 1–18.

Kim, K. K., Karunanayaka, P., Privitera, M. D., Holland, S. K., & Szaflarski, J. P. (2011). Semantic association investigated with functional MRI and independent component analysis. *Epilepsy & Behavior*, *20*(4), 613–622.

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

Kiran, S., Meier, E. L., Kapse, K. J., & Glynn, P. A. (2015). Changes in task-based effective connectivity in language networks following rehabilitation in post-stroke patients with aphasia. *Frontiers in Human Neuroscience*, *9*, 316.

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

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, *63*(8), 1279–1292.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

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.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Mbwana, J., Berl, M. M., Ritzl, E. K., Rosenberger, L., Mayo, J., Weinstein, S., … Gaillard, W. D. (2009). Limitations to plasticity of language network reorganization in localization related epilepsy. *Brain*, *132*(2), 347–356.

Meinzer, M., Beeson, P. M., Cappa, S., Crinion, J., Kiran, S., Saur, D., … Thompson, C. K. … Neuroimaging in Aphasia Treatment Research Workshop. (2013). Neuroimaging in aphasia treatment research: Consensus and practical guidelines for data analysis. *NeuroImage*, *73*, 215–224.

Meltzer, J. A., Postman-Caucheteux, W. A., McArdle, J. J., & Braun, A. R. (2009). Strategies for longitudinal neuroimaging studies of overt language production. *NeuroImage*, *47*(2), 745–755.

Nadeau, S. E. (2014). Neuroplastic mechanisms of language recovery after stroke. In J. I. Tracy, B. M. Hampstead, & K. Sathian (Eds.), *Cognitive plasticity in neurologic disorders* (pp. 61–84). Oxford: Oxford University Press.

Naeser, M. A., Helm-Estabrooks, N., Haas, G., Auerbach, S., & Srinivasan, M. (1987). Relationship between lesion extent in 'Wernicke's area' on computed tomographic scan and predicting recovery of comprehension in Wernicke's aphasia. *Archives of Neurology*, *44*(1), 73–82.

Nelson, D., McEvoy, C., & Schreiber, T. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://w3.usf.edu/FreeAssociation

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Pillay, S. B., Stengel, B. C., Humphries, C., Book, D. S., & Binder, J. R. (2014). Cerebral localization of impaired phonological retrieval during rhyme judgment. *Annals of Neurology*, *76*(5), 738–746.

Postman-Caucheteux, W. A., Birn, R. M., Pursley, R. H., Butman, J. A., Solomon, J. M., Picchioni, D., … Braun, A. R. (2010). Single-trial fMRI shows contralesional activity linked to overt naming errors in chronic aphasic patients. *Journal of Cognitive Neuroscience*, *22*(6), 1299–1318.

Price, C. J., & Crinion, J. (2005). The latest on functional imaging studies of aphasic stroke. *Current Opinion in Neurology*, *18*(4), 429–434.

Price, C. J., Crinion, J., & Friston, K. J. (2006). Design and analysis of fMRI studies with neurologically impaired patients. *Journal of Magnetic Resonance Imaging*, *23*(6), 816–826.

Rau, S., Fesl, G., Bruhns, P., Havel, P., Braun, B., Tonn, J.-C., & Ilmberger, J. (2007). Reproducibility of activations in Broca area with two language tasks: A functional MR imaging study. *American Journal of Neuroradiology*, *28*(7), 1346–1353.

Reppen, R., Ide, N., & Suderman, K. (2005). *American National Corpus (ANC) second release*. Philadelphia: Linguistic Data Consortium.

Rice, G. E., Hoffman, P., & Lambon Ralph, M. A. (2015). Graded specialization within and between the anterior temporal lobes. *Annals of the New York Academy of Sciences*, *1359*(1), 84–97.

Robson, H., Zahn, R., Keidel, J. L., Binney, R. J., Sage, K., & Lambon Ralph, M. A. (2014). The anterior temporal lobes support residual comprehension in Wernicke's aphasia. *Brain*, *137*(3), 931–943.

Rombouts, S. A., Barkhof, F., Hoogenraad, F. G., Sprenger, M., Valk, J., & Scheltens, P. (1997). Test-retest analysis with functional MR of the activated area in the human visual cortex. *American Journal of Neuroradiology*, *18*(7), 1317–1322.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*(2), 217–236.

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

Sanai, N., Mirzadeh, Z., & Berger, M. S. (2008). Functional outcome after language mapping for glioma resection. *New England Journal of Medicine*, *358*(1), 18–27.

Saur, D., & Hartwigsen, G. (2012). Neurobiology of language recovery after stroke: Lessons from neuroimaging studies. *Archives of Physical Medicine and Rehabilitation*, *93*(1), S15–S25.

Saur, D., Lange, R., Baumgaertner, A., Schraknepper, V., Willmes, K., Rijntjes, M., & Weiller, C. (2006). Dynamics of language reorganization after stroke. *Brain*, *129*(6), 1371–1384.

Seghier, M. L., Kherif, F., Josse, G., & Price, C. J. (2011). Regional and hemispheric determinants of language laterality: Implications for preoperative fMRI. *Human Brain Mapping*, *32*(10), 1602–1614.

Seghier, M. L., Lazeyras, F., Pegna, A. J., Annoni, J.-M., Zimine, I., Mayer, E., … Khateb, A. (2004). Variability of fMRI activation during a phonological and semantic language task in healthy subjects. *Human Brain Mapping*, *23*(3), 140–155.

Selnes, O. A., Knopman, D. S., Niccum, N., Rubens, A. B., & Larson, D. (1983). Computed tomographic scan correlates of auditory comprehension deficits in aphasia: A prospective recovery study. *Annals of Neurology*, *13*(5), 558–566.

Selnes, O. A., Niccum, N., Knopman, D. S., & Rubens, A. B. (1984). Recovery of single word comprehension: CT-scan correlates. *Brain and Language*, *21*(1), 72–84.

Shah, P. P., Szaflarski, J. P., Allendorfer, J., & Hamilton, R. H. (2013). Induction of neuroplasticity and recovery in post-stroke aphasia by non-invasive brain stimulation. *Frontiers in Human Neuroscience*, *7*, 888.

Sharp, D. J., Scott, S. K., & Wise, R. J. S. (2004). Retrieving meaning after temporal lobe infarction: The role of the basal language area. *Annals of Neurology*, *56*(6), 836–846.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology*, *6*(2), 174–215.

Springer, J. A., Binder, J. R., Hammeke, T. A., Swanson, S. J., Frost, J. A., Bellgowan, P. S. F., … Mueller, W. M. (1999). Language dominance in neurologically normal and epilepsy subjects: A functional MRI study. *Brain*, *122*(11), 2033–2046.

Swinburn, K., Porter, G., & Howard, D. (2004). *Comprehensive Aphasia Test*. Hove: Psychology Press.

Szaflarski, J. P., Allendorfer, J. B., Banks, C., Vannest, J., & Holland, S. K. (2013). Recovered vs. not-recovered from post-stroke aphasia: The contributions from the dominant and non-dominant hemispheres. *Restorative Neurology and Neuroscience*, *31*(4), 347–360.

Szaflarski, J. P., Holland, S. K., Jacola, L. M., Lindsell, C., Privitera, M. D., & Szaflarski, M. (2008). Comprehensive presurgical functional MRI language evaluation in adult patients with epilepsy. *Epilepsy & Behavior*, *12*(1), 74–83.

Thompson, C. K., & den Ouden, D.-B. (2008). Neuroimaging and recovery of language in aphasia. *Current Neurology and Neuroscience Reports*, *8*(6), 475–483.

Turkeltaub, P. E., Messing, S., Norise, C., & Hamilton, R. H. (2011). Are networks for residual language function and recovery consistent across aphasic patients? *Neurology*, *76*(20), 1726–1734.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., ... Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289.

Tzourio-Mazoyer, N., Petit, L., Razafimandimby, A., Crivello, F., Zago, L., Jobard, G., ... Mazoyer, B. (2010). Left hemisphere lateralization for language in right-handers is controlled in part by familial sinistrality, manual preference strength, and head size. *Journal of Neuroscience*, *30*(40), 13314–13318.

van Oers, C. A. M. M., Vink, M., van Zandvoort, M. J. E., van der Worp, H. B., de Haan, E. H. F., Kappelle, L. J., ... Dijkhuizen, R. M. (2010). Contribution of the left and right inferior frontal gyrus in recovery from aphasia: A functional MRI study in stroke patients with preserved hemodynamic responsiveness. *NeuroImage*, *49*(1), 885–893.

Vannest, J. J., Karunanayaka, P. R., Altaye, M., Schmithorst, V. J., Plante, E. M., Eaton, K. J., ... Holland, S. K. (2009). Comparison of fMRI data from passive listening and active-response story processing tasks in children. *Journal of Magnetic Resonance Imaging*, *29*(4), 971–976.

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

Warren, J. E., Crinion, J. T., Lambon Ralph, M. A., & Wise, R. J. S. (2009). Anterior temporal lobe connectivity correlates with functional outcome after aphasic stroke. *Brain*, *132*(12), 3428–3442.

Weiller, C., Isensee, C., Rijntjes, M., Huber, W., Müller, S., Bier, D., ... Diener, H. C. (1995). Recovery from Wernicke's aphasia: A positron emission tomographic study. *Annals of Neurology*, *37*(6), 723–732.

Wild, C. J., Davis, M. H., & Johnsrude, I. S. (2012). Human auditory cortex is sensitive to the perceived clarity of speech. *NeuroImage*, *60*(2), 1490–1502.

Wilke, M., & Lidzba, K. (2007). LI-tool: A new toolbox to assess lateralization in functional MR-data. *Journal of Neuroscience Methods*, *163*(1), 128–136.

Wilson, S. M., Bautista, A., Yen, M., Lauderdale, S., & Eriksson, D. K. (2017). Validity and reliability of four language mapping paradigms. *NeuroImage: Clinical*, *16*, 399–408.

Wilson, S. M., Eriksson, D. K., Schneck, S. M., & Lucanie, J. M. (2018). A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLoS One*, *13*(2), e0192773.

Woermann, F. G., Jokeit, H., Luerding, R., Freitag, H., Schulz, R., Guertler, S., ... Ebner, A. (2003). Language lateralization by Wada test and fMRI in 100 patients with epilepsy. *Neurology*, *61*(5), 699–701.

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

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, *4*(1), 58–73.

Yagata, S. A., Yen, M., McCarron, A., Bautista, A., Lamair-Orosco, G., & Wilson, S. M. (2017). Rapid recovery from aphasia after infarction of Wernicke's area. *Aphasiology*, *31*(8), 951–980.

You, X., Adjouadi, M., Guillen, M. R., Ayala, M., Barreto, A., Rishe, N., ... Gaillard, W. D. (2011). Sub-patterns of language network reorganization in pediatric localization related epilepsy: A multisite study. *Human Brain Mapping*, *32*(5), 784–799.

Zahn, R., Drews, E., Specht, K., Kemeny, S., Reith, W., Willmes, K., ... Huber, W. (2004). Recovery of semantic word processing in global aphasia: A functional MRI study. *Brain Research. Cognitive Brain Research*, *18*(3), 322–336.

Zijdenbos, A. P., Dawant, B. M., Margolin, R. A., & Palmer, A. C. (1994). Morphometric analysis of white matter lesions in MR images: Method and validation. *IEEE Transactions on Medical Imaging*, *13*(4), 716–724.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.