**Predicting online behavioural responses to transcranial direct current stimulation in
stroke patients with anomia**

Hope, T.M.H.^{1,2*}, Ondobaka S.¹, Akkad, H.¹, Nardo, N.³,
& Crinion, J.¹, PhD.

1. Institute of Cognitive Neuroscience, University College London, UK.
2. Wellcome Centre for Human Neuroimaging, University College London, UK
3. MRC Cognition and Brain Sciences Unit, University of Cambridge, UK
4. Institute of Health and Wellbeing, University of Glasgow, UK

*Corresponding author; E-mail: t.hope@ucl.ac.uk

Short title: Predicting neurostimulation responses in anomia

Corresponding author: Thomas Hope, t.hope@ucl.ac.uk, Institute of Cognitive Neuroscience, 17-19 Queen Square, London WC1N 3AR.

References: 36

Words: (abstract): 279; (body): 6,654

Abstract

BACKGROUND: Anomia, or difficulty naming common objects, is the most common, acquired impairment of language. Effective therapeutic interventions for anomia typically involve massed practice at high doses, requiring significant investment from patients and therapists. Aphasia researchers have increasingly looked to neurostimulation to accelerate these treatment effects, but the evidence behind this intervention is sparse and inconsistent.

AIM: Here, we hypothesised that group-level neurostimulation effects might belie more systematic structure at the individual level. We sought to test the hypothesis by attempting to predict the immediate (online), individual-level behavioural effect of neurostimulation in patients with anomia.

METHODS: 36 stroke patients, each with anomia at least 6 months post-onset, were asked to perform naming and judgement tasks both with concurrent neurostimulation and with sham stimulation. Using clinical, (pre-stimulation) behavioural and MRI data, and Partial Least Squares regression, we attempted to predict the effect of neurostimulation on accuracies and reaction times in both tasks. Model performance was assessed via cross-validation, and performances compared to that of a null model, which predicted the mean neurostimulation effects for all patients.

RESULTS: Models derived from pre-stimulation data consistently out-performed the null model when predicting neurostimulation effects on accuracies and reaction times in both judgement and naming. Notably, we could predict declines in performance just as well as improvements.

CONCLUSIONS: Inter-patient variation in online responses to neurostimulation is to some extent systematic and predictable. That declines in performance were just as predictable as improvements, implies that the effect of neurostimulation on performance in patients with anomia is unlikely to be a placebo. However, the online effect of the intervention appears to be just as likely to interfere with task performance, as it is to improve performance.

1. Introduction

Anomia, which primarily presents as a difficulty with naming common objects in pictures, is perhaps the most common symptom of acquired language impairment [1]. Treatments for anomia typically depend on massed practice, and require high doses to work [2]: to really benefit, patients and therapists need to invest significant time and effort. To try to reduce that effort, research interest has increasingly turned to potential accelerants of treatment for aphasia, including drugs [3] and neurostimulation [4]. One promising example of the latter type of intervention is transcranial Direct Current Stimulation (TDCS), which is both relatively inexpensive and tolerable by most patients [5].

A typical TDCS setup involves two electrodes: an active electrode and a return electrode. The active electrode is placed on the scalp over the target brain region, and the return electrode is placed elsewhere, either on the head (e.g. over the contralateral supraorbital area) or on a region away from the head (such as the contralateral shoulder). The TDCS electrodes interact to create an electrical field that alters cell membrane potentials, which in turn modulates the likelihood of neuronal firing in (and potentially around / beyond [Hunter 2013]) the target brain region [6]. Quite what kind of modulation it makes, is still a matter of debate. In broad terms, cathodal TDCS is thought to decrease the excitability of neurons in the target region, while anodal TDCS increases their excitability. However, this dichotomy is derived from studies stimulating the motor cortices [7], leaving the precise effects on other brain regions largely unknown [8]. And even when stimulating the motor cortices, those precise effects also depend on the detailed set-up (e.g. the size of the electrodes, the distance between them, and the intensity of the stimulation) [9].

Nevertheless, there is increasing evidence that tDCS can be used to modulate language task performance, in both in the neurologically normal [5, 10, 11] and the damaged [4, 6, 9, 12] brain. In the study of anomia, one natural stimulation site is the left inferior frontal gyrus (LIFG), both because activity in this region (which includes Broca's area) has a long association with semantic language [13], and because the same activity is known to correlate with naming performance in stroke patients with aphasia [14-17]. Consistent with these findings, anodal TDCS has been shown to enhance stroke patients' naming accuracy when applied over the LIFG [18]. However, the evidence here is inconsistent because other studies have either failed to find the effect, or found it but only with cathodal tDCS (e.g. [19]).

One reason for this inconsistency might be that even when significant effects are found in this domain, they tend to be modest [4, 12]: i.e. these small effects might either be artefacts of noise in those experiments where they appear, or they might be real, but hard to distinguish because of the noise in these studies. In either case, one might justifiably doubt the clinical relevance of these effects. However, even in studies that fail to find these effects, at least some individual patients typically appear to respond very well indeed. The primary aim of this study was to test the claim that those responses are really just artefacts of noise.

Following the model of our own recent work in the study of therapy effects [20], we use prediction to test the claim. If individual, task-related, behavioural responses to neurostimulation are indeed driven by noise (either measurement noise or stimulation-induced noise), then there should be no particular structure to the individual variance in neurostimulation responses: i.e. they should not be consistent or predictable. However, if we can find that structure, and show that it is consistent by using it to predict new patients' responses, those responses cannot be pure artefacts of noise. From this perspective, apparently modest group-level effects might belie more complex individual-level effects, in which some subset of those patients tested are much more likely to respond, perhaps much more significantly, than the group-level literature suggests. To the extent that these online effects predict longer-term gains [18], the implication is that we might better stratify patients for TDCS-enhanced treatment by identifying those who are most likely to benefit.

2. Methods

2.1 Participants

The study included 36 participants, all of whom had survived a first-time left-hemisphere stroke at least one year before entering the study, had normal hearing, normal or corrected-to-normal visual acuity and no previous history of neurological or psychiatric disease, as well as no contraindications to MRI scanning. Inclusion criteria were: (i) anomia as determined by the naming subtest of the Comprehensive Aphasia Test [21]; (ii) good single word comprehension as assessed by the spoken words comprehension subtest of the Comprehensive Aphasia Test; (iii) relatively spared ability to repeat single monosyllabic words from the Psycholinguistic Assessments of Language Processing in Aphasia [22]; (iv)

absence of speech apraxia as determined by the Apraxia Battery for Adults [23]. Participants were excluded if they had any contraindications for scanning, had more than one stroke or had any other significant neurological or psychiatric conditions. Informed consent was obtained from all participants prior to participation under approval from Central London Research Ethics Committee, UK.

All participants' lesions were assessed, by JC, as either sparing or damaging the LIFG. Participants who had suffered extensive damage to LIFG were assigned to the Broca's area Damaged (BD) group; otherwise, patients were assigned to the Broca's area Preserved (BP) group. It was assumed that BD patients would not respond well to neurostimulation applied to the lesioned cortex ; following prior evidence of right homologue adaptation during recovery, in patients with LIFG damage (e.g. [24]), we chose to stimulate the right homologue of LIFG (i.e. RIFG) in these patients.

Anodal stimulation was applied at each site, following the logic of the Bimodal Balance Recovery (BBR) model [25]. Initially advanced as a model of recovery from post-stroke hemiparesis, this model builds on prior findings suggesting that the bilateral motor cortices exhibit mutual inhibition [26]. In that context, the natural assumption might be that *inhibitory* neurostimulation to the contralateral motor cortex, in stroke, might disinhibit the ipsilateral motor cortex, facilitating recovery: this is sometimes called 'inter-hemispheric rebalancing' [27]. The BRR model posits an exception to this rule, when the ipsilateral motor cortex has suffered extensive damage, extinguishing or severely reducing the 'structural reserve' available to make use of inter-hemispheric rebalancing. In this case, rather than inhibiting the contralateral hemisphere, we might best promote recovery by recruiting those contralateral homologues [28]. The patients in our BD group had extensive damage to LIFG, so we employed anodal (putatively excitatory) rather than cathodal (putatively inhibitory) neurostimulation to its contralateral homologue. Our hypothesis was that this would yield comparable responses to neurostimulation in the LIFG, for those patients whose LIFG was largely preserved.

2.2 Experimental Design

Participants completed two simultaneous fMRI / TDCS sessions, one with anodal stimulation and the other with sham stimulation, separated by a one-week wash-out period. In each session, participants completed both a picture naming task and a size judgement task:

stimulation responses were measured as their accuracies or reaction times on each task during anodal stimulation minus accuracies / reaction times on the same task during sham stimulation. Each task also manipulated both visual object information (by adding noise to the images) and auditory object information (with phonemic cues), but our stimulation response variables average across these conditions.

2.3 Behavioural tasks

The Naming task required participants to name each presented object as quickly and as accurately as possible. The Judgment task required them to decide whether the object would fit inside a microwave oven (yes/no verbal responses), the aim being to study lexical retrieval in naming while controlling for object identification, decision making, and a vocal speech response.

Each object picture was a black and white line drawing of a concrete object. A trial started with a 1s presentation of a fixation cross, followed by simultaneous appearance of the object picture with the visual and auditory cues (squiggly lines / geometric shapes superimposed on the images, and auditory stimuli corresponding to the object's name or its first phoneme). Pictures were displayed for 2.5s, followed by a blank screen for 0.35s (see Figure 1). The order of experimental conditions was pseudo-randomized within a block (i.e., avoiding more than three trials of the same condition in a row), and objects were counterbalanced across participants to control for psycholinguistic features like frequency, concreteness, imageability, initial phoneme. Experimental trials were interspersed with fixation-only rest periods, to optimise the sampling of the BOLD response [29]. And the inter-trial interval was set to 3.85s, to vary the spatiotemporal synchrony between the trial structure and the image acquisition.

Stimulus presentation and behavioural data acquisition were controlled using Matlab. Overt spoken responses were recorded online using a dual-channel, noise-cancelling fibre optical microphone system (FOMRI III; <http://www.optoacoustics.com>). These recordings were reviewed offline to calculate speech accuracy scores (% correct) for each participant. Auditory cues were delivered via MR-compatible headphones (MR Confon, Magdeburg, Germany; www.mr-confon.de). Participants completed a short training period before entering the scanner for each session, both to familiarise them the task and to practice speaking softly,

thus minimising motion in the scanner. Stimuli used during the training were not used during the fMRI session.

2.4 Stimuli

A total of 480 target words was selected from the IPNP database [30] (n=220), and from the MRC Psycholinguistic Database [31] (n=260). Object names were monosyllabic words with a consonant-vowel-consonant (CVC) phonological structure. Phonemic cues were digitally recorded (at 44.1 kHz) from a male native English speaker in a soundproof room, then cropped at the offset of the vowel to form the initial phoneme cue (e.g., /bɒ/ for 'box').

2.5 TDCS and concurrent fMRI

TDCS was delivered during the fMRI experiment by using an MR-compatible stimulation system (neuroConn; https://www.neurocaregroup.com/dc_stimulator_mr.html) via a pair of MR-compatible leads and rectangular rubber electrodes (5x7 cm), allowing for a current density of 0.057 mA/cm² (cf. [11]). For those participants identified (by JC, from structural MRI images) as 'LIFG intact', the anode was placed over the LIFC (equivalent to position FC5 in 10-20 EEG nomenclature; cf. Figure 1D), and the cathode placed over the contralateral frontopolar cortex (FP2). For participants identified as 'LIFG destroyed', these stimulation sites were mirrored with the anode placed over the Right hemisphere IFC (RIFC). Both electrodes and the sites on the scalp where the electrodes were placed were covered with EEG conductive paste to ensure a flush and comfortable fit between the electrode surface and the scalp. Electrodes were secured to the head using 3M Coban elastic wrap bandage and placed in adherence with the manufacturer's MR safety guidelines. Care was taken in connecting the leads backward along the centre of the scanner bore to minimize the possibility of radio frequency-induced heating, and to ensure that any gradient switching-induced AC currents were well below the level that might cause stimulation. The stimulator was placed outside the Faraday cage of the scanner, and the stimulating current was fed to the participant through two stages of radio frequency filtration to prevent interference being picked up by the scanner.

A scanner pulse triggered the onset of the stimulation at a given slice in the acquisition sequence. The current was increased slowly during the first 15 sec to the desired stimulation

threshold (2 mA), termed the ramp-up phase. A constant direct current (2 mA) was delivered for 20 min. At the end of the stimulation period, the current was decreased to 0 mA over 1 sec (ramp-down). For sham stimulation, the ramp-up phase was followed by 15 sec of 2 mA stimulation, which was immediately followed by a 1 sec ramp-down phase. This active sham protocol makes for a more efficient double-blinding process: both stimulation and sham protocols produced sensations of comparable quality (a mild tingling, typically under the electrode placed over the contralateral orbital/frontopolar edge). Participants habituated to the sensation quickly and reported minimal discomfort with no adverse sensations during both anodal and sham TDCS stimulation runs. The position of the anode and cathode electrodes for each subject was recorded and reproduced across scanning sessions.

2.6 MRI acquisition and analysis

Whole-brain imaging was performed on a 3T Siemens TIM-Trio system (Siemens, Erlangen, Germany) at the Wellcome Centre for Human Neuroimaging. T2*-weighted echo-planar images (EPI) with BOLD contrast were acquired using a 12-channel head coil. Structural MRI data were processed into binary lesion images using the Automatic Lesion Identification toolbox [32], running under SPM 12 in Matlab. We employed the (120,231) left-hemisphere voxels from the “fuzzy” lesion images, produced by the toolbox, to represent patients’ lesions. Voxel values in these regions range from 0 to 1, with higher values indicated greater evidence that the voxel is damaged, and evidence is derived by comparing tissue intensity in each voxel to intensities from a population of neurologically normal controls.

2.7 Data analysis

2.7.1 Behavioural data

Response accuracy in the judgement task (i.e. deciding whether an object would fit inside a microwave oven) was defined relative to a normative standard, derived from a separate experiment run using Amazon’s web-based Mechanical Turk platform (MTurk; see [33]). 32 adult participants viewed all of the 480 stimuli that we used in the main experiment (without visual or auditory cues) and used yes / no button press responses to indicate whether each object would fit in a microwave oven. 62/480 items did not reach our threshold of 70% agreement in these responses, so were excluded from the main analysis. Judgement and naming reaction times were timed from stimulus onset until the patient responded. Stimulation responses for both tasks were calculated as the naming accuracy / reaction time

during anodal stimulation minus the naming accuracy / reaction time during sham stimulation.

2.7.2 Modelling Methods

Our key aim here was to assess whether individual patients' stimulation responses could be predicted from pre-stimulation data alone. The pre-stimulation data included: (a) clinical data (age at stroke onset, sex, time post-stroke, pre- and post-stroke handedness, years of formal schooling and total lesion volume); (b) behavioural data, derived from a wide-ranging battery of standardised tests of post-stroke impairment severity; and (c) lesion load variables, derived from pre-stimulation structural MRI, as previously described.

The behavioural variables were derived from a comprehensive battery of language and non-language tests, administered to assess participants' language and cognitive abilities. The tests included: the (full) Comprehensive Aphasia Test [21], the PALPA 8 and 9 [22], the Boston Naming Test [34], the Pyramids and Palm Trees test [35], the Cattell Culture Fair IQ Test (Scale 2, Form A) [36], the Rey-Osterrieth Complex Figure Test [37], digit-span tasks from the Wechsler Adult Intelligence Scale [38], the Delis-Kaplan Executive Function System test [39], the Hopkins Verbal Learning Test [40], and the Children's Sustained Attention to Response Task [41]. A complete list of all of the 96 variables derived from these tests is included in Supplementary Material. And the lesion data consisted of lesion probability values from a total of 120,213 2mm³ voxels from the left hemisphere of the Montreal Neurological Institute brain. Given the very large number of predictors relative to the sample size ($n = 36$), we employed Partial Least Squares (PLS) regression to reduce the predictors' dimensionality in an efficient manner.

2.7.3 Predictive models

We used PLS regression, as implemented in Matlab 2019a. The behavioural model employed 96 predictors, recording scores on our battery of pre-treatment language and cognitive assessments (as described in detail in [42]). The lesion data were encoded as described previously, into 398 lesion load variables: however, all patients had left-hemisphere lesions, and in fact all patients had zero lesion load in 220/398 regions. These were removed from the analysis (leaving 178 variables), but their removal had no substantive effect on the results.

We trained models employing predictors derived from each data type separately, and all higher order combinations of data types.

2.7.4 Model assessment and model comparison

Predictive performance was assessed with cross-validation. We report results using 1,000 times 10-fold cross-validation here, but analyses employing different types of cross-validation were substantially similar. Absolute measures of predictive performance are suspect in small samples, so we assessed our models in relative terms, by comparing them to an empty, or baseline (i.e. null) model, which simply predicts the average treatment response for all patients in the group. This model reflects our null hypothesis, that treatment responses are *not predictable at the individual level*, leaving group-level averages as the only recourse. When our empirical models outperform the empty model, we reject the null hypothesis, concluding that individual treatment responses are predictable, at least to some extent. We compare models by recording the Mean Squared Error (MSE) for predictions made in each of 1,000 repetitions of a 10-fold cross-validation. These folds are kept identical across models, so the MSE values can be compared pair-wise.

However, traditional paired tests are not appropriate on their own in this case because different partitions of the data will create overlapping training datasets, which are therefore not independent of each other. Accordingly, while we use the traditional, paired, non-parametric Wilcoxon signed rank test to compare MSEs across models, we further threshold those statistics with paired permutation test [x]. The test construes the two vectors to be compared as having labels, reflecting the models used to generate them. If null hypothesis is that those labels are arbitrary, because the models' performance do not differ except by chance. We therefore create a null distribution of paired test statistics by randomly permuting those labels *within each pair* and repeating the original paired (signed rank) test. If the original statistic is extreme relative to the null distribution, we conclude that the performance difference between the models is significant ($p < 0.05$) after a correction for FamilyWise Error (FWE).

2.7.5 Model interpretation

PLS regression models can be interpreted by examining the weights of each of their components on each of the original variables. However, this approach can be challenging when there are multiple components to consider, and because the sign of each component is

arbitrary: i.e. positive weights on a given component do not necessarily imply a positive relationship between the highly weighted independent variables and the dependent variable(s). We circumvent these issues with ‘data perturbation’.

The data perturbation procedure involves permuting random subsamples of the empirical independent variables and recording the effect of the perturbation on the model’s predictions. The PLS model beta weights themselves are fixed based on the original empirical data: our goal is not to fit further models, but rather to better understand the relationships that have already been encoded. We do this by: (a) permuting random subsets of the independent variables; (b) observing the effect on the models’ predictions; and (c) relating perturbed variable values to the resulting predictions with mass univariate correlation analyses. The resultant correlation coefficients approximate the influence that each variable has on the model’s predictions. We ran 1,000 iterations of the process per model, yielding a total sample size of 18 (patients) * 1,000 (iterations), including both perturbed independent variables and the resultant, predicted dependent variable (treatment response). Repeated analyses with this number of iterations yielded very consistent coefficients for all of the models we report: across ten repetitions of 1,000 iterations of this process, all pairwise correlations between derived weights on behavioural and lesion variables were >0.99.

One complicating factor in this analysis, is that the variable valence of both predictor and response variables varies. For example, one of our behavioural variables was scores on the Comprehensive Aphasia Test’s auditory word comprehension task [21]; in this case, higher scores imply better performance. However, other variables encode the numbers of errors made in a given task, such as phonemic errors in naming, and in this case higher scores imply worse performance. Moreover, remembering that participants were asked to respond as accurately and quickly as possible in the judgement and naming tasks, larger accuracy rates and smaller reaction times each imply better performance. Since our data perturbation method assigns positive or negative weights to each variable, these shifts in the valence of those variables will complicate our results. For the sake of parsimony, we therefore distinguish between ‘intuitive’ and ‘counter-intuitive’ weights. For lesion variables, intuitive weights predict better stimulation responses with less damage (and vice versa), where ‘better’ means either higher accuracies or faster responses. Counter-intuitive weights imply the opposite association, between more damage and better stimulation responses. Intuitive weights on behavioural variables associate better pre-stimulation performance with better

stimulation responses (and vice versa). This terminology hopefully allows for a clearer discussion of the results.

3. Results

3.1 Patient characteristics

Our sample of 36 stroke patients included 26 men and 10 women. The mean age was 59.1yrs (± 12.5 yrs), the mean time post-stroke at assessment was 10.4yrs (± 8.0 yrs), and the patients' mean years of schooling was 13.8yrs (± 2.4 yrs). The sample was divided into two sub-groups, depending on whether their lesions had preserved or partially preserved the left inferior frontal gyrus (Broca's area Preserved: BP; or Broca's area Destroyed: BD). The two groups were not significantly different in age at stroke onset, time post-stroke at assessment, or years of schooling (all $p > 0.1$). The mean lesion size across the whole sample was 125.0cm^3 ($\pm 106.4\text{cm}^3$), with patients in the BD group suffering significantly larger lesions ($66.7\text{cm}^3 \pm 106.4\text{cm}^3$) than the BP group ($183.3\text{cm}^3 \pm 115.1\text{cm}^3$): $Z = 3.33$, $p < 0.001$. See Figure 1 for a lesion frequency plot.

Figure 1. Lesion frequency plot, for the sample of 36 patients

3.2 In-scanner behaviour and stimulation responses

Table 1 reports participants' in-scanner performance under both sham and anodal stimulation. Under sham stimulation, the BP group were significantly more accurate than the BD group in naming ($Z = 2.48$, $p = 0.013$) but not judgement ($Z = 1.43$, $p = 0.150$). The BP group were also numerically faster than the BD group in the judgement task, though this did not reach

significance ($Z = 1.91$, $p = 0.057$), and there was also no significant difference in reaction times for the naming task ($Z = 0.21$, $p = 0.837$). Under anodal stimulation, the results were similar in that the BP group were more accurate in naming than the BD group ($Z = 2.86$, $p = 0.004$), and there were no other significant differences between the groups (all $p > 0.3$).

		ALL	BP	BD
SHAM	Judgement %	74.9 (± 15.2)	78.2 (± 13.7)	71.6 (± 16.4)
	Judgement RT (s)	1.63 (± 0.28)	1.54 (± 0.31)	1.72 (± 0.24)
	Naming %	66.8 (± 16.9)	72.2 (± 16.6)	61.3 (± 15.9)
	Naming RT (s)	1.57 (± 0.26)	1.57 (± 0.23)	1.57 (± 0.29)
ANODAL	Judgement %	75.5 (± 13.8)	77.0 (± 13.3)	73.9 (± 14.6)
	Judgement RT (s)	1.64 (± 0.31)	1.59 (± 0.34)	1.70 (± 0.27)
	Naming %	67.8 (± 16.0)	74.7 (± 14.2)	60.8 (± 14.8)
	Naming RT (s)	1.55 (± 0.23)	1.58 (± 0.24)	1.52 (± 0.22)

Table1: In-scanner performance. Accuracies and reaction times (mean and standard deviation) for the whole sample and both sub-groups, in the naming and judgement tasks respectively, under both sham and anodal stimulation.

Crucially, there were no significant effects of neurostimulation, either on accuracies or on reaction times in either task, either across the group as a whole or in either of the two sub-groups: all $p > 0.2$. This is despite significant variance at the individual level in all four independent variables: i.e. the stimulation appears to significantly alter some patients' accuracies and / or reaction times, even though there is no consistent, group-level effect. We hypothesised the null effects at the group level might belie more systematic structure in the patients' individual responses to neurostimulation, which are illustrated in Figure 2. We reasoned that, if the changes themselves are predictable, these individual-level effects cannot be dismissed as artefacts of noise.

Figure 2: Stimulation responses.

Mean accuracies (top) and reaction times (bottom), in sham vs. anodal stimulation, in the judgement (left) and naming (right) tasks. The numbers of patients whose stimulation responses were positive (/ negative) in each case were:

judgement accuracies: 15 (21);

naming accuracies: 20 (16);

judgement reaction times (measured in seconds): 20 (16); naming reaction times: 19 (17).

3.3 Predictive performance

With 3 data types to consider (clinical, behavioural and lesion data), we employed a total of 8 model configurations: i.e. 7 combinations of predictor data types, plus an 8th combination for the null model, which does not employ any predictor data (because it simply predicts the mean stimulation response of the patients in its training sample). Table 2 records the results, highlighting the best models for stimulation responses (accuracies and reaction times) in the judgement and naming tasks respectively.

DATA	JUDGEMENT %	NAMING %	JUDGEMENT RT	NAMING RT
NULL	1.034/0.033	1.028/0.027	1.033/0.033	1.032/0.030
DEM	1.213/0.075	0.984/0.057	1.315/0.090	1.379/0.115
BEH	1.146/0.081	0.838/0.091	0.861/0.128	0.980/0.093
LES	0.846/0.120	0.780/0.095	0.828/0.073	<u>0.916/0.080</u>
DEM+BEH	1.153/0.078	0.775/0.09	<u>0.723/0.118</u>	1.020/0.040
DEM+LES	0.848/0.121	0.779/0.094	0.830/0.073	0.917/0.08
BEH+LES	<u>0.842/0.107</u>	0.749/0.087	0.819/0.067	0.919/0.08
DEM+BEH+LES	0.845/0.107	<u>0.748/0.087</u>	0.821/0.070	0.933/0.066

Table 2: Model predictive performances. Median and inter-quartile ranges of Mean Squared Errors (MSE) achieved by models driven by different combinations of data, across 1,000 repetitions of a 10-fold cross-validation process. Folds are kept consistent across models. The best performing models, with the lowest median MSE, employed behavioural and lesion variables for the judgement task, and clinical, behavioural and lesion variables for the judgement task. DEM = clinical data; BEH – behavioural data; LES = lesion data. Notably, all stimulus response variables were standardised here prior to learning, so MSEs are comparable across them.

All of our dependent variables – stimulation responses for accuracies and reaction times in the naming and judgement tasks – appeared to be predictable. Each could be predicted by some combination of pre-stimulation data, at better (lower) error rates than those associated with the null model. Lesion data were a component of the best models' predictors for 3/4 of the dependent variables. The exception was for predicting stimulation effects on reaction times in the judgement task, where the best model included clinical and behavioural data only (see Table 2); however, lesions alone were also able to make better-than-null-model predictions for this variable.

The best models for each response variable were also all significantly better (after FWE correction) than the next best, implying that their advantages, though numerically small, were highly consistent across cross-validation folds. The best models were also all very significantly better than the null model: all FWE $p < 0.001$. These results imply that the small advantages enjoyed by these models are very consistent across the 1,000 repetitions of the cross-validation process. Since all stimulation response variables were standardised prior to running the predictive analyses, their MSEs are also at least somewhat comparable. Stimulation effects on naming accuracy appeared more predictable than stimulation effects on naming reaction times, whereas this order was reversed for the judgement task.

Finally, no model relied simply on predicting improvements (increasing accuracies or decreasing reaction times) consequent to neurostimulation. If any model is really only predicting change in one direction, it follows that the predictions for those patients whose performance changed in the other direction should be less accurate. To test this, we employed a Bayesian analysis of variance [43], measuring whether the direction of empirical change (stimulation responses greater versus lesser than sham responses) mediated the MSEs of each

of our models' mean predictions. In fact, there was strong evidence *against* a mediating effect for all dependent variables: BF = 0.04 (judgement accuracy); 0.04 (judgement reaction times); 0.03 (naming accuracy); and 0.08 (naming reaction times). As in past work [44], we interpret these results as evidence of systematicity / predictability in both the improvements and the declines in performance, here associated with neurostimulation.

3.4 Examining the role of stimulation site

One feature of the experimental design deserves particular consideration: half of the patients in this sample were stimulated in the LIFG, whereas the other half, with no or minimal preservation of LIFG after stroke, were stimulated in the RIFG. If responses to stimulation at each site are different, our models might have simply discovered this feature of the experimental design, confounding our results. There were no significant differences in the stimulation responses for each group in either the judgement task (Wilcoxon rank sum test; accuracies: $z = 0.98$, $p = 0.33$; reaction times: $z = 1.12$, $p = 0.26$) or the naming task (accuracies: $z = 1.60$, $p = 0.10$; reaction times: $z = 0.97$, $p = 0.33$). However, this still leaves open the possibility that there is some subtler interaction, for example with lesion site, which is driving the whole-sample predictive results.

One way to test this would be to repeat the analyses described in the last section for each sub-group separately. However, this test is confounded by the significant reduction in the sample sizes of each sub-group relative to the whole group, which is already smaller than might be preferred for predictive modelling. Instead, we compared the performances of our best models for each task when predicting between stimulation sites, to their performance when predicting across stimulation sites.

If stimulation site is a key driver of the original whole-sample results, it should be particularly important to train our models with some examples of patients who were stimulated at LIFG, *and* some examples of patients who were stimulated at RIFG. It follows that predictions for each sub-group based on training only with the other ('between-group splits'), should be less accurate than predictions derived based on training with exemplars from both sub-groups ('across-group splits'). We tested this by running each of our four best PLS regression models (one for each dependent variable) through a split-half cross-validation with the between-group split, and comparing its MSE to a distribution of MSEs obtained for

the same model over 1,000 cross-validation runs with across-group splits. To the extent that our results for each dependent variable are confounded by stimulation site, we expected more across-group MSEs to be smaller than the between-group MSE for that variable.

The proportions of across-group MSEs that were in fact smaller than each between-group MSE were: 84% (judgement accuracies), 15% (naming accuracies), 61%, (judgement reaction times) and 36% (naming reaction times).

Quite how high these proportions would need to be, to demonstrate a stimulation-site confound, remains an open question. A threshold of $> 95\%$ seems consistent with the standard definition of significance but might be considered too conservative here. Instead, we interpret these proportions as convincing evidence that stimulation site is not driving our results for reaction times in both tasks, and for accuracy in the naming task, because they imply that the between-group MSE was either easier than most across-group MSEs, or within the inter-quartile range of those across-group MSEs. Neurostimulation induced changes in judgement accuracy, however, were outside that range (84%), suggesting that stimulation-site confounds might be more relevant in this case.

3.5 Interpreting the best models

Each of the four best PLS regression models assigns non-zero weights to all of its predictors, and in this sense, none of the predictors are, strictly, redundant in any model. Nevertheless, while we do report all model weights, either in Figure 2 below (for lesion-related weights), or in supplementary material (for non-lesion weights), for the sake of parsimony we concentrate only on the most important variables here: i.e. those assigned the weights of the largest magnitude.

Figure 2: Weights on the lesion variables in our best models for stimulation responses for accuracies in the judgement task (top) and naming task (middle), and also for reaction times in the naming task (bottom). Note that no brain image is shown for the model predicting stimulation responses on reaction times in the judgement task, because lesion data played no part in the best model for that variable.

In those models that included lesion data, the strongest weights, implying that more damage drives less improvement / more decline in performance after neurostimulation, were assigned to the inferior temporal cortex (judgement accuracy), the putamen (naming accuracy), and the superior longitudinal fasciculus (SLF) (naming reaction times). Note that in the first two cases, these weights were negative, because higher accuracies imply better performance, but in the third case (naming reaction times), the weight was positive, implying

a greater tendency to lengthening reaction times with more damage to the SLF. In these models, the strongest weights implying the opposite relationship, with more lesion damage predicting worse responses to neurostimulation, were found in the cuneus (judgement accuracy: 0.53), and the cingulum (naming accuracy: 0.50; and naming reaction times; 0.50).

In those models that included behavioural data, the strongest weights predicting greater performance improvement with better pre-stimulation performance were on scores in the REYS immediate copy task (0.04; judgement accuracy); successful self-corrections in confrontation naming (0.05; naming accuracy), and the PALPA 9 repetition task (0.10; judgement reaction times). In the same models, the strongest weights implying the opposite relationship – predicting worse responses to neurostimulation with better pre-treatment performance – were scores on the CAT verbal fluency task (0.06; judgement accuracies), the CAT summary score for all auditory comprehension tasks (0.09; naming accuracy), and the recognition subtest of the Hodgkin Verbal Learning Test (0.23; judgement reaction times).

Finally, in those (two) models that employed clinical data, the strongest weights were on time post-stroke when predicting naming accuracies (-0.06), implying that patients assessed sooner after stroke responded better to the neurostimulation. When predicting judgement reaction times, the strongest weight was on total lesion volume (0.14), suggesting that patients with larger lesions tended to have longer reaction times during anodal than sham stimulation.

4. Discussion

TDCS has attracted growing attention, as a potential accelerant for the rehabilitation of aphasia [12], and evidence suggests that the immediate, online impact of tDCS might predict those longer-term rehabilitation effects [4]. However, the real impact of tDCS is still

somewhat unclear, because, while many studies report significant facilitation of language via tDCS (e.g. [4, 10, 19, 45]), these effects appear modest and inconsistent [6, 9, 12].

One potential explanation for this inconsistency follows from a recent study of the effects of tDCS in verbal fluency in the healthy (young) adult brain [46], in which the authors reported a ‘paradoxical’ decline in performance rather than the expected improvement. If these interference effects can also occur in the damaged brain, modest or null neurostimulation effects at the group level might belie a more complex dichotomy of improvements and declines at the individual level. Certainly, a visual inspection of our own patients’ neurostimulation responses is consistent with this notion, because some patients’ performance improves and others’ performance declines (see Figure 2).

Notably, this is also the expected pattern if neurostimulation responses were driven entirely by (measurement and / or stimulation-induced) noise. But noise-induced changes should not be predictable, and many of the models that we tested made significantly better predictions than those of a baseline model, in which each patient’s prediction was simply the mean stimulation response of the group in each task (see Table 1). We also found strong [47] evidence against a mediating effect of the direction of change, on the predictability of change: i.e. prediction errors for each of our four best models were not significantly different depending on whether the empirical change was positive or negative, and we employed a Bayesian analysis of variance [43] to quantify the evidence in favour of this null result. Since neurostimulation responses were significantly predictable, and both directions of response appeared equally predictable, it follows that neither direction of response is an artefact of noise. We hope that this result will encourage future reports of these studies to distinguish participants whose performance declines with tDCS, rather than simply distinguishing ‘responders’ from ‘non-responders’, as is currently common [12].

Notably, it was typically also just as easy to predict stimulation responses for one stimulation site (left inferior frontal gyrus) after training only with patients stimulated at the contralateral homologue (and vice versa), as it was when training with exemplars from both stimulation sites. As far as we know, this is the first evidence that the Bimodal Balance Recovery (BBR) model [25] might apply to recovery from aphasia, as well as motor impairment, after stroke: i.e., to some approximation, the functional role of the LIFG in patients whose strokes spared it, is comparable to that of the RIFG, in patients with severe damage to the LIFG.

Weights assigned by each of our four best models were broadly plausible. The emphasis on the inferior temporal cortex (ITC), in the best model for judgement accuracies, is consistent with the role that the ITC is thought to play in object recognition via the ventral visual stream [48, 49]. The emphasis on the putamen and supervening white matter, in the model for naming accuracies, is consistent with that region's extensive connectivity to left hemisphere language networks [50]. And the emphasis on the superior longitudinal fasciculus in the model for naming reaction times, is also consistent with studies relating the preservation of this tract to long-term language outcomes after stroke [51, 52]. Notably, lesion damage in each of the regions that were positively associated with neurostimulation responses (i.e. more damage = better response) was significantly anti-correlated with lesion damage in those regions where damage had the more intuitive association with neurostimulation responses (i.e. more damage = poorer response). This suggests that these models' more counter-intuitive weights (more damage = better response) might reflect the contingent distribution of the lesions suffered by these patients. In other words, regions where more damage appears to improve responses to neurostimulation might best be explained as regions where damage predicts the preservation of other regions required to benefit from the intervention.

Weights on behavioural variables were more difficult to interpret. The weight on the immediate-copy condition from the Rey Complex Figure Task, in the model for judgement accuracies, is sensible in that it implies that better recovered or more preserved executive functioning skills [53], after stroke, might encourage better responses to neurostimulation in the size judgement task. The weight on self-corrections of errors in confrontation naming, in the model for naming accuracies, is plausible in that it links better preserved / recovered error monitoring, in naming, to improvements in naming capacity consequent to neurostimulation. However, the emphasis on scores in the PALPA9 word repetition task, in the model for judgement reaction times (better scores = faster responses), is difficult to explain. And in all models, counter-intuitive weights (better pre-stimulation performance predicts worse stimulation responses) were stronger than intuitive weights. This perhaps suggests an element of regression to the mean, with patients suffering greater impairment more likely to respond better to neurostimulation, and vice versa.

However, given the relatively small sample size ($n = 36$), compared to the number of predictors that we have considered ($>100,000$ when lesion data are included), we would caution against strong interpretations of these weights. Our modelling approach offers a principled and (largely) unbiased way to measure whether stimulation responses are predictable in principal. This method is also demonstrably at least somewhat robust to the extremely high dimensionality of the data that we have considered. But we assume that the ‘best’ (or perhaps even ‘right’) models for these responses will include only a subset of these variables, and there is no guarantee that the weights on those variables will be consistent with what we see here. We have not yet found a principled way to identify that subset, which also preserves predictive power when properly nested within a cross-validation process. One attractive, if heuristic option would be to explore sparse variants of Partial Least Squares regression (e.g. [54]). This is a focus of ongoing work.

Nevertheless, our results suggest that individual responses to neurostimulation in both judgement and naming tasks, are at least somewhat systematic and predictable. Stimulation-induced declines in performance appeared to be just as predictable as stimulation-induced improvements, encouraging a greater focus on ‘decliners’ in future neurostimulation studies, in addition to ‘responders’ and (possibly) ‘non-responders’. And the apparent ease with which we could predict stimulation responses with the between-site split adds confidence that responses to stimulation at each site were consistent – or at least consistently related to patients’ behavioural impairments and lesions. We hope that these results will encourage a more personalised approach to the application of neurostimulation in post-stroke aphasia, maximising its potential to improve the efficiency of therapies for post-stroke aphasia.

ACKNOWLEDGMENTS

This study was supported by Wellcome (203147/Z/16/Z; 106161/Z/14/Z), and the Stroke Association (TSA PDF 2017/02). The funders had no participation in the design and results of this study.

DATA

The data described in this study is available to accredited researchers from JC, on request.

602 REFERENCES

- 603 1. Laine, M. and N. Martin, *Anomia: Theoretical and clinical aspects*. 2013: Psychology Press.
- 604 2. Brady, M.C., et al., *Speech and language therapy for aphasia following stroke*. Cochrane
605 database of systematic reviews, 2016(6).
- 606 3. Bakheit, A.M., *Drug treatment of poststroke aphasia*. Expert Rev Neurother, 2004. **4**(2): p.
607 211-7.
- 608 4. de Aguiar, V., C.L. Paolazzi, and G. Miceli, *tDCS in post-stroke aphasia: the role of stimulation
609 parameters, behavioral treatment and patient characteristics*. Cortex, 2015. **63**: p. 296-316.
- 610 5. Nitsche, M.A., et al., *Transcranial direct current stimulation: state of the art 2008*. Brain
611 stimulation, 2008. **1**(3): p. 206-223.
- 612 6. Alharbi, M.F., S. Armijo-Olivo, and E.S. Kim, *Transcranial direct current stimulation (tDCS) to
613 improve naming ability in post-stroke aphasia: A critical review*. Behavioural Brain Research,
614 2017. **332**: p. 7-15.
- 615 7. Giordano, J., et al., *Mechanisms and Effects of Transcranial Direct Current Stimulation*. Dose-
616 Response, 2017. **15**(1): p. 1559325816685467.
- 617 8. Nitsche, M.A. and W. Paulus, *Excitability changes induced in the human motor cortex by
618 weak transcranial direct current stimulation*. The Journal of physiology, 2000. **527 Pt 3**(Pt 3):
619 p. 633-639.
- 620 9. Ulanov, M.A., Y.Y. Shtyrov, and T.A. Stroganova, *Transcranial Direct Current Stimulation as a
621 Tool to Induce Language Recovery in Patients with Post-Stroke Aphasia*. Neuroscience and
622 Behavioral Physiology, 2019. **49**(9): p. 1169-1180.
- 623 10. Meinzer, M., et al., *Transcranial direct current stimulation over multiple days improves
624 learning and maintenance of a novel vocabulary*. Cortex, 2014. **50**: p. 137-147.
- 625 11. Holland, R., et al., *Speech facilitation by left inferior frontal cortex stimulation*. Current
626 Biology, 2011. **21**(16): p. 1403-1407.
- 627 12. Biou, E., et al., *Transcranial direct current stimulation in post-stroke aphasia rehabilitation: A
628 systematic review*. Annals of Physical and Rehabilitation Medicine, 2019. **62**(2): p. 104-121.
- 629 13. Broca, P., *Remarques sur le siège de la faculté du langage articulé, suivies d'une observation
630 d'aphémie (perte de la parole)*. Bulletin de la Société Anatomique, 1861. **6**: p. 330-357.
- 631 14. Crinion, J.T. and A.P. Leff, *Recovery and treatment of aphasia after stroke: functional
632 imaging studies*. Curr Opin Neurol, 2007. **20**(6): p. 667-73.
- 633 15. Crosson, B., et al., *Functional MRI of language in aphasia: a review of the literature and the
634 methodological challenges*. Neuropsychol Rev, 2007. **17**(2): p. 157-77.
- 635 16. Postman-Caucheteux, W.A., et al., *Single-trial fMRI shows contralesional activity linked to
636 overt naming errors in chronic aphasic patients*. J Cogn Neurosci, 2010. **22**(6): p. 1299-318.
- 637 17. Fridriksson, J., et al., *Activity in preserved left hemisphere regions predicts anomia severity in
638 aphasia*. Cereb Cortex, 2010. **20**(5): p. 1013-9.
- 639 18. Baker Julie, M., C. Rorden, and J. Fridriksson, *Using Transcranial Direct-Current Stimulation
640 to Treat Stroke Patients With Aphasia*. Stroke, 2010. **41**(6): p. 1229-1236.
- 641 19. Monti, A., et al., *Improved naming after transcranial direct current stimulation in aphasia*. J
642 Neurol Neurosurg Psychiatry, 2008. **79**(4): p. 451-3.
- 643 20. Aguilar, O.M., et al., *Lesion-site-dependent responses to therapy after aphasic stroke*. Journal
644 of Neurology, Neurosurgery & Psychiatry, 2018.
- 645 21. Swinburn, K., Porter, G., and Howard, D., *Comprehensive Aphasia Test*. 2004: Psychology
646 Press.
- 647 22. Kay, J., R. Lesser, and M. Coltheart, *Psycholinguistic assessments of language processing in
648 aphasia (PALPA)*. Hove, UK: Laurence Erlbaum Associates, 1992.
- 649 23. Dabul, *Apraxia battery for adults (second ed.)*. 2000, Austin, Texas: Pro-Ed.
- 650 24. Rosen, H., et al., *Neural correlates of recovery from aphasia after damage to left inferior
651 frontal cortex*. Neurology, 2000. **55**(12): p. 1883-1894.

25. Di Pino, G., et al., *Modulation of brain plasticity in stroke: a novel model for neurorehabilitation*. Nature Reviews Neurology, 2014. **10**(10): p. 597.
26. Duque, J., et al., *Intermanual differences in movement-related interhemispheric inhibition*. Journal of cognitive neuroscience, 2007. **19**(2): p. 204-213.
27. Harris-Love, M.L. and R.M. Harrington, *Non-Invasive Brain Stimulation to Enhance Upper Limb Motor Practice Poststroke: A Model for Selection of Cortical Site*. Frontiers in Neurology, 2017. **8**(224).
28. Pundik, S., et al., *Recovery of post stroke proximal arm function, driven by complex neuroplastic bilateral brain activation patterns and predicted by baseline motor dysfunction severity*. Frontiers in human neuroscience, 2015. **9**: p. 394.
29. Henson, R., *Efficient experimental design for fMRI*. Statistical parametric mapping: The analysis of functional brain images: p. 193-210.
30. Szekely, A., et al., *A new on-line resource for psycholinguistic studies*. Journal of memory and language, 2004. **51**(2): p. 247-250.
31. Wilson, M., *MRC psycholinguistic database: Machine-usable dictionary, version 2.00*. Behavior research methods, instruments, & computers, 1988. **20**(1): p. 6-10.
32. Seghier, M.L., et al., *Lesion identification using unified segmentation-normalisation models and fuzzy clustering*. Neuroimage, 2008. **41**(4): p. 1253-66.
33. Buhrmester, M.D., S. Talaifar, and S.D. Gosling, *An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use*. Perspectives on Psychological Science, 2018. **13**(2): p. 149-154.
34. Kaplan, E., H. Goodglass, and S. Weintraub, *Boston naming test*. 2001: Pro-ed.
35. Howard, D. and K.E. Patterson, *The Pyramids and Palm Trees Test: A test of semantic access from words and pictures*. 1992: Thames Valley Test Company.
36. Cattell, R.B. and A. Cattell, *Cattell Culture Fair Intelligence Test: A Measure of "g"*. 1960: Bobbs-Merrill.
37. Osterrieth, P.A., *Le test de copie d'une figure complexe; contribution a l'etude de la perception et de la memoire*. Archives de psychologie, 1944.
38. Wechsler, D., *Wechsler adult intelligence scale—Fourth Edition (WAIS—IV)*. San Antonio, Texas: Psychological Corporation, 2014.
39. Delis, D.C., E. Kaplan, and J.H. Kramer, *Delis-Kaplan executive function system*. 2001.
40. Brandt, J. and R.H. Benedict, *Hopkins verbal learning test--revised: professional manual*. 2001: Psychological Assessment Resources.
41. Robertson, I.H., et al., *Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects*. Neuropsychologia, 1997. **35**(6): p. 747-758.
42. Nardo, D., et al., *Less is more: neural mechanisms underlying anomia treatment in chronic aphasic patients*. Brain, 2017. **140**(11): p. 3039-3054.
43. Wetzels, R., R.P.P.P. Grasman, and E.-J. Wagenmakers, *A Default Bayesian Hypothesis Test for ANOVA Designs*. The American Statistician, 2012. **66**(2): p. 104-111.
44. Hope, T.M.H., et al., *Right hemisphere structural adaptation and changing language skills years after left hemisphere stroke*. Brain, 2017.
45. Holland, R. and J. Crinion, *Can tDCS enhance treatment of aphasia after stroke?* Aphasiology, 2012. **26**(9): p. 1169-1191.
46. Klaus, J. and G. Hartwigsen, *Failure to Improve Verbal Fluency with Transcranial Direct Current Stimulation*. Neuroscience, 2020. **449**: p. 123-133.
47. Jeffreys, H., *Theory of Probability*. 1961, Oxford, UK: Oxford University Press.
48. Cichy, R.M., et al., *Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence*. Scientific Reports, 2016. **6**: p. 27755.
49. DiCarlo, James J., D. Zoccolan, and Nicole C. Rust, *How Does the Brain Solve Visual Object Recognition?* Neuron, 2012. **73**(3): p. 415-434.

50. Viñas-Guasch, N. and Y.J. Wu, *The role of the putamen in language: a meta-analytic connectivity modeling study*. Brain Struct Funct, 2017. **222**(9): p. 3991-4004.
51. Dick, A.S. and P. Tremblay, *Beyond the arcuate fasciculus: consensus and controversy in the connectional anatomy of language*. Brain, 2012. **135**(12): p. 3529-3550.
52. Ramsey, L.E., et al., *Behavioural clusters and predictors of performance during recovery from stroke*. Nature human behaviour, 2017. **1**: p. 0038.
53. Smith, S.R. and N.E. Zahka, *Relationship between accuracy and organizational approach on the Rey-Osterrieth Complex Figure and the Differential Ability Scales: A pilot investigation*. Child Neuropsychology, 2006. **12**(6): p. 383-390.
54. Chun, H. and S. Keleş, *Sparse partial least squares regression for simultaneous dimension reduction and variable selection*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2010. **72**(1): p. 3-25.

717 **Supplementary Material**

718 **A. Behavioural variables**

719 In what follows, we list all 96 behavioural variables that are considered in the main analyses.

- 720 1. Total score in all CAT comprehension tasks for auditory stimuli
- 721 2. Total score in all CAT comprehension tasks for visual stimuli
- 722 3. Total score in all CAT auditory repetition tasks
- 723 4. Total score in all CAT picture naming tasks
- 724 5. CAT semantic memory task
- 725 6. Total score in all CAT verbal fluency tasks
- 726 7. CAT category fluency task
- 727 8. CAT letter fluency task
- 728 9. CAT recognition memory task
- 729 10. CAT object use gesture task
- 730 11. CAT arithmetic task
- 731 12. CAT task on the comprehension of spoken words
- 732 13. CAT task on the comprehension of written words
- 733 14. CAT task on the comprehension of spoken sentences
- 734 15. CAT task on the comprehension of written sentences
- 735 16. CAT task on the comprehension of spoken paragraphs
- 736 17. CAT single-word auditory repetition task
- 737 18. Performance when repeating single-syllable words (subset of the above)
- 738 19. CAT complex word repetition task
- 739 20. CAT non-word repetition task
- 740 21. CAT digit span task
- 741 22. CAT sentence repetition task
- 742 23. CAT object (picture) naming task
- 743 24. CAT action (picture) naming task
- 744 25. CAT scene (picture) description task
- 745 26. Appropriate words used in the scene description task
- 746 27. Inappropriate words used in the scene description task
- 747 28. CAT word reading task
- 748 29. CAT complex word reading task
- 749 30. CAT function word reading task
- 750 31. CAT non-word reading task
- 751 32. CAT written scene (picture) description
- 752 33. CAT writing to dictation task
- 753 34. PALPA8 non-word repetition task
- 754 35. PALPA8 non-word reading task
- 755 36. PALPA9 word repetition (High Imageability High Frequency)
- 756 37. PALPA9 word repetition (High Imageability Low Frequency)
- 757 38. PALPA9 word repetition (Low Imageability High Frequency)
- 758 39. PALPA9 word repetition (Low Imageability Low Frequency)
- 759 40. PALPA9 non-word repetition task
- 760 41. Boston Naming Test (cued and uncued)
- 761 42. Boston Naming Test (cued)

- 762 43. Boston Naming Test (uncued)
763 44. Confrontation Naming task (exactly correct responses)
764 45. Confrontation Naming task (response is a synonym of correct response)
765 46. Confrontation Naming task (errors self-corrected to correct responses)
766 47. Confrontation Naming task (multiple word answers containing correct responses)
767 48. Confrontation Naming task (total correct responses)
768 49. Confrontation Naming task total semantic errors
769 50. Confrontation Naming task total phonemic errors
770 51. Confrontation Naming task total omission errors
771 52. Confrontation Naming task miscellaneous errors
772 53. Confrontation Naming task coordinate errors (response is a different item belonging to
773 the same category as the target word).
774 54. Confrontation Naming task superordinate errors (response is the target word's
775 superordinate category).
776 55. Confrontation Naming task superordinate errors (response is a more specific word to the
777 target and does not match the picture).
778 56. Confrontation Naming task associate errors. Response is semantically associated to the
779 target word (noun, verb or adjective) and does not fit into any of the other semantic error
780 categories. This includes sound effects that demonstrate semantic knowledge of the target
781 word.
782 57. Confrontation Naming task part-whole errors. Response is the correct for a part of the
783 whole picture. Alternatively, a response is correct for the whole picture, rather than the
784 target part.
785 58. Confrontation Naming task circumlocution errors. Response is a meaningful, multiword
786 description of the target without producing the target word.
787 59. Confrontation Naming task formal phonemic errors. Response with at least 50%
788 phonemes overlapping with the target word in any position to create a real word.
789 60. Confrontation Naming task phonemic errors. Response with at least 50% phonemes
790 overlapping with the target word (or an accepted synonym) in any position to create a
791 non-word.
792 61. Confrontation Naming task partial phonemic errors. Response has at least one phoneme
793 that is related or unrelated to the target
794 62. Confrontation Naming task neologism errors. Response is a nonword that does not fulfil
795 the above phonemic error criteria.
796 63. Confrontation Naming task formal omission (incorrect) errors. Response that does not fit
797 into any other category, does not have any semantic relationship to the target. This also
798 includes stereotyped responses.
799 64. Confrontation Naming task omission (missing) errors. No response is given
800 65. Confrontation Naming task perseveration errors. Perseverated response repeated from a
801 recent previous target.
802 66. Confrontation Naming task mixed errors. Response contains a mix of semantic and
803 phonemic errors
804 67. Confrontation Naming task miscellaneous (unrelated) errors. Response is a real word, but
805 semantically or phonologically unrelated to the target word.
806 68. Confrontation Naming task self-wrong errors. Response is initially correct but changed to
807 an incorrect response
808 69. Confrontation Naming task visual errors. Response that is visually similar to the target
809 but is not semantically related.
810 70. Pyramids and Palm Trees Test
811 71. Cattell Culture Fair IQ Test, Scale 2, Form A

- 812 72. RBM
- 813 73. Rey-Osterrieth Complex Figure Test, immediate copy condition
- 814 74. Rey-Osterrieth Complex Figure Test, 3-minute delay copy condition
- 815 75. Rey-Osterrieth Complex Figure Test, 30-minute delay copy condition
- 816 76. Wechsler Adult Intelligence Scale digit span total (scaled)
- 817 77. Wechsler Adult Intelligence Scale digit span total
- 818 78. Wechsler Adult Intelligence Scale digit span (forwards)
- 819 79. Wechsler Adult Intelligence Scale digit span (backwards)
- 820 80. Hopkins Verbal Learning Test total recall score
- 821 81. Hopkins Verbal Learning Test delayed recall score
- 822 82. Hopkins Verbal Learning Test recognition score
- 823 83. Children's Sustained Attention to Response Task, percent hits
- 824 84. Children's Sustained Attention to Response Task, omissions
- 825 85. Children's Sustained Attention to Response Task, reaction time
- 826 86. Children's Sustained Attention to Response Task, errors
- 827 87. Delis-Kaplan Executive Function System test, scanning task
- 828 88. Delis-Kaplan Executive Function System test, numbers task
- 829 89. Delis-Kaplan Executive Function System test, letters task
- 830 90. Delis-Kaplan Executive Function System test, switching task
- 831 91. Delis-Kaplan Executive Function System test, correct sorts
- 832 92. Delis-Kaplan Executive Function System test, free description
- 833 93. Delis-Kaplan Executive Function System test, recognition and description
- 834 94. Delis-Kaplan Executive Function System test, combined description
- 835 95. Delis-Kaplan Executive Function System test, verbal score
- 836 96. Delis-Kaplan Executive Function System test, perceptual score

837

838 **B. Non-lesion weights in the best model for each dependent variable**

839 The best model for reaction times in naming employed only lesion data (see Table 2). In what
840 follows, we report weights assigned via data perturbation for the non-lesion variables in the
841 best models for the other three dependent variables that we considered. The first table
842 concerns weights on clinical variables (i.e. demographic variables and total lesion volume).
843 The next table concerns behavioural variables, and is presented in the same order as the list of
844 variable descriptions in the last section.

845

846 Table S1: Weights on clinical variables where they appear in the best model for a given
847 stimulation response variable.

VARIABLE	NAME %	JUDGE RT
Sex	-0.025	-0.006
Age	-0.048	-0.017
Schooling years	-0.007	0.017
Handedness before stroke	-0.026	-0.011
Handedness after stroke	0.014	-0.095
Time post stroke	-0.064	0.043
Lesion volume	-0.018	0.141

848

849 Table S2: Weights on behavioural variables where they appear in the best model for a given
850 stimulation response variable.

TASK	JUDGE %	NAME %	JUDGE RT
CAT_AUDCOMP_TOTAL_of66	-0.025	-0.092	-0.022
CAT_VISCOMP_TOTAL_of62	-0.048	0.005	-0.150
CAT_REPETITION_TOTAL_of74	-0.007	-0.028	0.008
CAT_NAMING_TOTAL_of94	-0.026	-0.029	-0.051
CAT_semanticmemory	0.014	-0.053	-0.134
CAT_verbalfluency_total	-0.064	-0.073	-0.138
CAT_verbalfluency_animals	-0.018	0.006	-0.048
CAT_verbalfluency_S	-0.092	-0.021	-0.173
CAT_recognitionmemory	0.005	0.012	0.020
CAT_gestureobjectuse	-0.028	0.004	-0.162
CAT_arithmetic	-0.029	-0.010	0.011
CAT_comp_spokenwords	-0.053	0.017	-0.133
CAT_comp_writenwords	-0.073	-0.008	-0.204
CAT_comp_spokensentences	0.006	0.016	0.051
CAT_comp_writtensentences	-0.021	0.015	-0.066
CAT_comp_spokenparagraphs	0.012	0.014	-0.003
CAT_wordrepetition	0.004	0.023	0.030
CAT_wordrepetition_1syl	-0.010	-0.017	-0.026
CAT_complexwordrepetition	0.017	0.013	0.070
CAT_nonwordrepetition	-0.008	-0.002	-0.015
CAT_digitstringrepetition	0.016	0.002	-0.004
CAT_sentencerepetition	0.015	0.013	0.020
CAT_namingobjects	0.014	-0.017	0.105
CAT_namingactions	0.023	-0.025	-0.023
CAT_spokenpicturedescription_total	-0.017	-0.011	-0.003
CAT_spokenpicturedescription_appropriateICWS	0.013	-0.030	0.017

CAT_spokenpicturedescription_inappropriateICWS	-0.002	0.020	-0.127
CAT_reading_words	0.002	-0.007	0.009
CAT_reading_complexwords	0.013	-0.005	0.051
CAT_reading_functionwords	-0.017	-0.038	-0.014
CAT_reading_nonwords	-0.025	-0.005	-0.101
CAT_writing_picturenames	-0.011	0.023	-0.124
CAT_writing_todictation	-0.030	0.013	-0.127
PALPA8_nonword_repetition	0.020	-0.015	0.057
PALPA8_nonword_reading	-0.007	-0.003	-0.043
PALPA9_wordrepetition_HIHF	-0.005	0.006	0.042
PALPA9_wordrepetition_HILF	-0.038	-0.016	-0.027
PALPA9_wordrepetition_LIHF	-0.005	0.024	0.108
PALPA9_wordrepetition_LILF	0.023	-0.047	0.071
PALPA9_nonwordrepetition	0.013	0.020	0.102
BNT_spontaneousandstim	-0.015	-0.009	-0.031
BNT_allcues	-0.003	0.007	0.001
BNT_nocues	0.006	0.006	-0.038
CN_correct_CR	-0.016	-0.001	-0.024
CN_correct_SY	0.024	0.017	0.106
CN_correct_SC	-0.047	0.053	-0.170
CN_correct_MR	0.020	0.020	0.080
CN_correct_TOTAL	-0.009	0.014	-0.047
CN_semantic_CE	0.007	-0.005	0.149
CN_semantic_SUP	0.006	-0.003	0.078
CN_semantic_SUB	-0.001	-0.035	0.079
CN_semantic_AE	0.017	-0.020	0.113
CN_semantic_PW	0.053	-0.020	0.149
CN_semantic_CL	0.020	0.016	0.104
CN_semantic_TOTAL	0.014	-0.025	0.164
CN_phonemic_FP	-0.005	0.026	-0.080
CN_phonemic_PE	-0.003	0.004	-0.027
CN_phonemic_PH	-0.035	-0.014	-0.141
CN_phonemic_NE	-0.020	-0.008	-0.122
CN_phonemic_TOTAL	-0.020	0.013	-0.134
CN_omission_IR	0.016	0.012	0.027
CN_omission_NR	-0.025	-0.006	-0.008
CN_omission_TOTAL	0.026	-0.010	0.017
CN_misc_PS	0.004	-0.045	0.028
CN_misc_MX	-0.014	-0.028	0.047
CN_misc_UR	-0.008	0.042	-0.022
CN_misc_SW	0.013	-0.029	-0.052
CN_misc_VE	0.012	-0.020	0.113
CN_misc_TOTAL	-0.006	0.023	-0.006
PPT	-0.010	0.032	-0.134
CATTELS	-0.045	0.013	-0.078
RPM	-0.028	0.000	-0.047
REYS_immediatecopy	0.042	-0.019	0.030
REYS_3mindelaycopy	-0.029	-0.030	-0.157

REYS_30mindelaycopy	-0.020	-0.043	-0.161
WAIS_digitspan_scaledscore	0.023	0.026	-0.015
WAIS_digitspan_total	0.032	-0.038	0.016
WAIS_digitspan_forward	0.013	0.027	0.028
WAIS_digitspan_backwards	0.000	-0.044	-0.016
HVLT_totalrecall	-0.019	-0.049	-0.074
HVLT_delayedrecall	-0.030	-0.015	-0.064
HVLT_recognition	-0.043	0.013	-0.239
CSART_percenthits	0.026	-0.040	0.047
CSART_omissions	-0.038	-0.025	-0.011
CSART_RT	0.027	-0.043	0.050
CSART_errors	-0.044	-0.042	-0.005
DKEFS_TM_scanning	-0.049	-0.048	-0.172
DKEFS_TM_numbers	-0.015	-0.006	-0.119
DKEFS_TM_letters	0.013	-0.035	-0.068
DKEFS_TM_switching	-0.040	0.013	-0.098
DKEFS_CS_correctsorts	-0.025	-0.010	-0.172
DKEFS_CS_freedescription	-0.043	0.002	-0.189
DKEFS_CS_recognitiondescription	-0.042	0.006	-0.213
DKEFS_CS_combineddescription	-0.048	0.027	-0.223
DKEFS_CS_verbal	-0.006	-0.008	-0.135
DKEFS_CS_perceptual	-0.035	0.010	-0.216

851

852