Check for updates

RESEARCH ARTICLE

UPDATE **Measurement of language laterality using functional transcranial Doppler ultrasound: a comparison of different tasks [version 3; peer review: 3 approved, 1 approved with reservations]**

Zoe V.J. Woodhead iD, Holly A. Rutherford, Dorothy V.M. Bishop iD

Department of Experimental Psychology, University of Oxford, Oxford, UK

Abstract

Background: Relative blood flow in the two middle cerebral arteries can be measured using functional transcranial Doppler sonography (fTCD) to give an index of lateralisation as participants perform a specific task. Language laterality has mostly been studied with fTCD using a word generation task, but it is not clear whether this is optimal.

Methods: Using fTCD, we evaluated a sentence generation task that has shown good reliability and strong left lateralisation in fMRI. We interleaved trials of word generation, sentence generation and list generation and assessed agreement of these tasks in 31 participants (29 right-handers).

Results: Although word generation and sentence generation both gave robust left-lateralisation, lateralisation was significantly stronger for sentence generation. Bland-Altman analysis showed that these two methods were not equivalent. The comparison list generation task was not systematically lateralised, but nevertheless laterality indices (LIs) from this task were significantly correlated with the other two tasks. Subtracting list generation LI from sentence generation LI did not affect the strength of the laterality index.

Conclusions: This was a pre-registered methodological study designed to explore novel approaches to optimising measurement of language lateralisation using fTCD. It confirmed that sentence generation gives robust left lateralisation in most people, but is not equivalent to the classic word generation task. Although list generation does not show left-lateralisation at the group level, the LI on this task was correlated with left-lateralised tasks. This suggests that word and sentence generation involve adding a constant directional bias to an underlying continuum of laterality that is reliable in individuals but not biased in either direction. In future research we suggest that consistency of laterality across tasks might have more functional significance than strength or direction of laterality on any one task.

Open Peer Review

Reviewer Status ✓ ✓ ✓ ?

	Invited Reviewers			
	1	**2**	**3**	**4**
version 3 (update) 11 Feb 2020			✓ report	? report
version 2 (revision) 15 Oct 2018				
version 1 24 Aug 2018	✓ report	✓ report		

1 **Ervin Sejdić** iD , University of Pittsburgh, Pittsburgh, USA

2 **Eva Gutierrez-Sigut** iD , University of Valencia , Valencia, Spain

3 **Guy Vingerhoets** iD , Ghent University, Ghent, Belgium

4 **Andreas Jansen** iD , Philipps University of Marburg, Marburg, Germany

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Dorothy V.M. Bishop (dorothy.bishop@psy.ox.ac.uk)

Author roles: Woodhead ZVJ: Data Curation, Formal Analysis, Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; **Rutherford HA**: Formal Analysis, Investigation, Writing – Original Draft Preparation; **Bishop DVM**: Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Background

It is well-established that language is predominantly processed in the left cerebral hemisphere in the majority of people. The functional significance of this bias is not understood, and it has been difficult to study the impact of individual variation because, on the one hand, those with atypical lateralisation are relatively rare (estimated as around 5 per cent of right-handers, Mazoyer et al., 2014), and on the other hand, assessment of cerebral lateralisation in non-clinical populations is time-consuming and challenging with large samples.

Measuring lateralisation with Functional Transcranial Doppler ultrasonography (fTCD)

Functional transcranial Doppler ultrasonography (fTCD) is a minimally invasive technique which is portable and relatively inexpensive, making it possible to investigate hemispheric dominance in relation to typical and impaired language function across a wider population. This method measures bilateral blood flow in the middle cerebral arteries (MCA) – which may be regarded as a proxy of cerebral activity in regions supplied by these vessels due to neurometabolic coupling (Lohmann et al., 2006). Although the territory of the MCA is variable between individuals, it consistently includes the key language regions of Broca's and Wernicke's areas (Deppe et al., 2004). In addition to being mobile and inexpensive, fTCD is largely unaffected by head movements and muscle artefacts during speech (Gutierrez-Sigut et al., 2015), meaning large numbers of participants can be investigated – including children and those who are unsuitable for other recording methods. The principal drawback of fTCD is that, while it is reliable at detecting increased blood flow in one hemisphere during an activation task with good temporal resolution, it cannot be used to localise function within a hemisphere.

Language tasks use to assess language lateralisation

The word generation task (Knecht et al., 1998a; Knecht et al., 1998b) has been used as the gold standard method for assessing cerebral lateralisation using fTCD – with good concordance with fMRI (Deppe et al., 2000; Somers et al., 2011) and Wada technique (Knecht et al., 1998a; Wada & Rasmussen, 1960), and adequate test-retest reliability (Knecht et al., 1998a). Although other tasks have been shown to be sensitive and reliable at assessing cerebral lateralisation, we still know relatively little about the key characteristics of tasks that give reliable left lateralisation. Indeed, it has been proposed that language laterality is multifactorial, and may dissociate within an individual from one task to another (Gaillard et al., 2004; Stroobant et al., 2009; Tailby et al., 2017).

Use of comparison tasks in fTCD and fMRI

One difference between fTCD and fMRI is that the latter typically includes a comparison task against which a language activation task is assessed. (We use here the term 'comparison task' rather than the more customary 'baseline task', to avoid confusion, as the term 'baseline' is used in fTCD to refer to the rest period prior to the language activation period). Subtracting activation from a comparison task is used in fMRI to remove activation associated with incidental aspects of task processing. For instance, if the task involves describing a picture, then a comparison task may involve viewing, but not describing, a picture, with the goal of removing activation associated with picture-viewing. Choice of comparison task in fMRI can have a substantial impact on the laterality index (Bradshaw et al., 2017); for example, Binder and colleagues (Binder et al., 2008) observed that laterality indices were stronger when language tasks were contrasted with active comparison tasks. In fTCD, left and right channels are mathematically equated by subtracting the levels during a baseline rest period from the whole trial, but the impact of subtracting the timecourse of a comparison task has not yet been ascertained.

Mazoyer and colleagues (2014) conducted the largest study to date comparing language lateralisation with fMRI in left-versus right-handers. They used a sentence generation task in which on each trial the participant viewed a pictured scene (see Figure 1a for example) and had to generate a sentence to describe it. Participants were trained to produce a particular sentence type beginning with a subject (e.g. "the children"), followed by a description of the subject (e.g. "wearing aprons"), a verb (e.g. "cook") and ending with a detail about the action (e.g. "in the kitchen"). As a comparison task, participants were asked to recite an overlearned word sequence (months of the year), while viewing a scrambled picture (see Figure 1b). Both the sentence generation task and the comparison task involved viewing a visual stimulus while generating words, so after subtracting the activation associated with the latter task from the former, the activation would reflect the additional brain activity elicited by selecting lexical items and assembling them into a grammatical sentence.
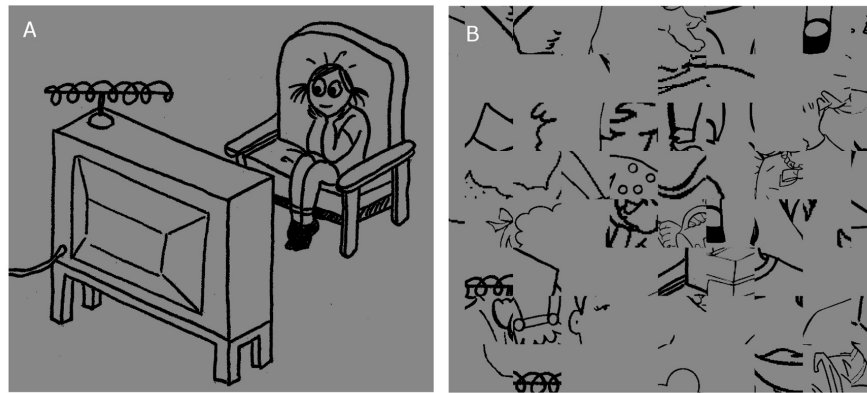
Figure 1. Examples of stimuli used in sentence generation and list generation tasks. A: An example of the line-drawing stimuli presented in the sentence generation tasks. **B**: An example of the scrambled line-drawing stimuli presented in the list generation task.

As will be discussed further below, because fTCD provides centisecond resolution for left-right blood flow differences, and the time course differs between tasks, it is not straightforward to simply subtract activation from a comparison task. Nevertheless, the potential benefits of having a comparison task to give a purer measure of activation during language generation are such that we decided to explore how feasible it would be to use a method in fTCD that was closely based on the approach of Mazoyer and colleagues. We predict that the effect of subtracting a comparison task would be the same as is observed in fMRI; i.e., it will yield higher LI values.

Our specific hypotheses were pre-registered on Open Science Framework (OSF) as follows:

Hypothesis 1: Individuals will be consistent in measures of hemispheric lateralization across different language domains; i.e., measures of lateralization (Lateralization index) will be statistically equivalent across word generation and sentence generation tasks.

Hypothesis 2: Lateralization indices acquired during the sentence generation task will be significantly larger when calculated relative to an active comparison task (number counting, i.e. automatic speech production) than when calculated relative to a resting baseline.

Methods
Deviation from pre-registered analysis plan
The analysis plan was pre-registered on OSF. There were three deviations from the pre-registered plan.

First, we have reworded hypothesis 2 to refer to an 'active comparison task' rather than an 'active baseline task', to avoid the confusion of 'baseline' being used with different meanings in fMRI and fTCD contexts (see above).

Second, data were analysed using custom scripts in R version 3.4.4 (R Core Team, 2017) rather than using the publicly-available software package, DopOSCCI, developed by

Badcock *et al*. (2012a) in Matlab. There were two reasons for this change: first, the R script had been developed in our group to fulfil the need for a reproducible and efficient method for processing large numbers of datasets, without using commercial (Matlab) software that required a licence (see Wilson & Bishop, 2018). As with DopOSCCI the analytic pipeline closely followed procedures developed by Deppe *et al*. (2004), with one additional option: the possibility of identifying brief periods of signal spiking or dropout and interpolating over these, to avoid rejecting trials. Wilson & Bishop (2018) compared results from DopOSCCI and the R script and found only small differences in the LIs computed by the two methods.

Third, 31 participants were tested instead of the planned group size of 30. The extra participant was recruited in case a dataset had to be excluded due to poor data quality, but ultimately all datasets were able to be used.

Participants
31 participants were recruited using flyers and advertisements distributed throughout the University of Oxford (11 males, age range 17–42, mean age 25 years). As our focus was normal range variation in language laterality, we did not use handedness as a selection criterion, but we asked participants to self-report their preferred hand for writing (left, right or ambidextrous): the sample included two left-handers. Participants were recruited after screening to confirm they had no known speech, language or learning impairment, or other diagnosed neurological disorder. In addition to these 31 participants, there were 4 participants who were recruited but for whom no data was collected: insonation of the MCA was not possible in two participants because a suitable temporal window could not be found; one participant became faint during the fTCD set-up and data collection was aborted; one participant disclosed a developmental reading delay after screening, which made them ineligible for participation.

Consent
Written informed consent for publication of anonymised data was obtained from the participants.

Power analysis

The required sample size to detect lateralisation greater than zero was calculated using a power analysis based upon effect sizes reported from previous studies using fTCD in comparable participant groups. A meta-analysis of 12 studies (Badcock *et al.*, 2012a; Bruckert, 2016; Dräger & Knecht, 2002; Grabitz *et al.*, 2016; Gutierrez-Sigut *et al.*, 2015; Illingworth & Bishop, 2009; Knecht *et al.*, 1998a; Krach & Hartje, 2006; Lust *et al.*, 2011; Somers *et al.*, 2011; Stroobant *et al.,* 2009; Whitehouse & Bishop, 2008) suggested that a sample size of 9 participants would be sufficient to detect the average effect size as found using the word generation task. However, as the expected strength of laterality with the sentence generation task is unknown, we planned to collect data from as many participants as possible within the time and resource constraints of the research team, with a goal of selecting a minimum of 30 participants.

Task design

All participants performed three tasks: the 'gold-standard' word generation (WG) task, sentence generation (SG) and list generation (LG). The task script and stimuli can be found at OSF (Woodhead *et al.*, 2018). The trials were presented using Psychophysics Toolbox version 3.0.13 (Brainard, 1997; Kleiner *et al.*, 2007) in MATLAB 2012b software (Mathworks Inc.), with 20 trials each of word, sentence and list generation interleaved in a pre-determined pseudorandomised order. Each trial had a common structure, as depicted in Figure 2. Participants were first asked to clear their mind, before the presentation of the task-specific stimulus for 2 seconds. This was followed by the presentation of a black cross in the centre of the screen for a duration of 10 seconds, indicating that participants should covertly generate the speech required for that trial type. After this period of covert generation, participants were prompted to "Report", and overtly state any speech they generated in the previous covert period. A period of covert speech generation allows measurement of changes in blood flow without the presence of any motor artefacts from speech production. Participants were then instructed to rest for 10 seconds. The rationale for the use of this particular structure was based upon a study by Gutierrez-Sigut and colleagues (2015), in which this same trial structure revealed strong effect sizes across both phonological and semantic fluency tasks, with a period of covert generation before overt report. Covert generation was shown to result in higher quality data in that fewer epochs were rejected from analysis.

Word generation task (WG). Following the structure presented in Figure 2, a single letter was presented as the stimulus on each trial of this task, indicating to the participant that they should generate as many words as possible beginning with that letter.

Sentence generation task (SG). The design was based on Mazoyer and colleagues (2014), though different stimulus materials were used to be more culturally appropriate for the UK. Participants were presented with a simple black line drawing on a grey background as the stimulus for this task. They were instructed to generate a sentence of a particular structure based on the image presented: beginning with a subject (e.g. "the children"), followed by a description of the subject (e.g. "wearing aprons"), a verb (e.g. "cook") and ending with a detail about the action (e.g. "in the kitchen"). Training before the task ensured that participants were comfortable with this sentence structure. An example of the line-drawing stimuli is depicted in Figure 1a.

List generation task (LG). LG acted as a comparison task, in which participants were instructed to repeatedly recite an over-learned word sequence, i.e., numbers one to ten. We picked this task after piloting showed that reciting months of the year (as used by Mazoyer *et al.*, 2014) was too demanding for many participants and led to poor behavioural performance. On list generation trials, participants were presented with a scrambled black line-drawing on the same grey background, to control for any possible perceptual effects, as depicted in Figure 1b.

Apparatus

fTCD data were acquired using a Doppler ultrasonography device (Doppler-Box™ X) with bilateral probes held in place using a DiaMon headset (DWL Elektronishe Systeme, Singen, Germany). Stimuli were presented on a desktop computer.

Procedure

After familiarisation with the stimuli with brief training exercises, the velocity of the blood flow in the left and right MCA was measured and recorded. A suitable, stable signal was isolated from each MCA before the start of the task, through the positioning of 2-MHz transducer probes, attached to a headset, bilaterally over each temporal window. The MCA was isolated at a depth of 45–65 mm with a velocity of 35–100cm/s, with its
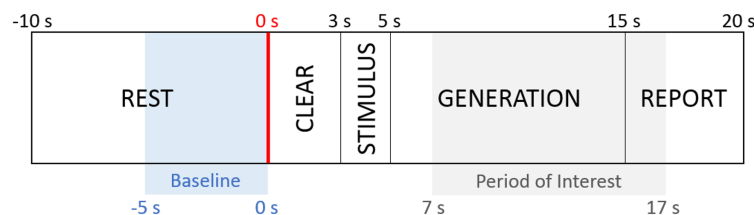

Figure 2. Schematic depicting the common structure of each trial. Schematic illustrating the timings used in each trial. The red line depicts the point at which the stimulus appears, while the period of interest is shown in grey, from 7 to 17 seconds. The blue area depicts the 5 second resting baseline period.

identity confirmed by increasing the signal depth until there was a bifurcation of the signal corresponding to the anterior communicating artery. Trials were split into two separate runs, each containing 30 trials, with a short break in between runs. Behavioural performance was recorded and the number of words spoken in each condition was noted. Trials in which participants failed to respond, or where the signal was lost from the probe, were marked to be excluded from analysis – this affected 4.65% of all trials.

fTCD Analysis

A custom R script (WSLG_Analysis.R, available from OSF (Bishop *et al.*, 2018b)) was used to obtain a laterality index (LI) from the raw cerebral blood flow velocity (CBFV) data from left and right channels.

The following steps are involved in data processing. Raw data are first down-sampled from 100 Hz to 25 Hz. The data are then segmented into epochs of 30 seconds, beginning 5 seconds before the presentation of the trigger stimulus for each trial. Before further processing, the script displays data for each trial, enabling the user to identify cases where there is a brief spiking or loss of signal: the signal was interpolated if the signal had an extreme value (quantile value < .0001 or >.9999) affecting just one time point; trials with longer periods of signal loss were excluded.

The next step is normalisation such that the values for blood flow velocity become independent from the angle of insonation (the angle at which the Doppler probe observes the flow of blood through the MCA) and the diameter of the MCA. Heart cycle integration then removes rhythmic modulations in blood flow velocity. Each epoch was baseline corrected using the baseline interval set within the 5-second rest period before trigger presentation for the resting baseline condition (Figure 2). Artefacts were identified as values below 60% and above 140% of mean CBFV and excluded from later analysis.

For each task, the left hemisphere CBFV minus right hemisphere CBFV is computed (L-R difference wave), and the point of maximal difference (the peak) within a predefined period of interest is identified. The period of interest began four seconds after stimulus presentation, and continued for 10 seconds, ending two seconds into the overt report period (see Figure 2). This is compatible with the time-course of the change in CBF in response to an external stimulus, which is thought to peak around 2–3 seconds after stimulus presentation (Gagnon *et al.*, 2015; Payne, 2018). The Lateralisation Index (LI) was calculated by taking the mean of the L-R difference wave in a two-second time window centred at that peak.

The 95% confidence interval of the LI from individual trials was used to classify participants as left-lateralised if their average LI was significantly greater than zero, and as right-lateralised if the LI was significantly less than zero. Participants were classified as bilateral if confidence interval of the LI spanned zero.

Statistical analysis

Statistical analyses were also performed in R version 3.4.4 (R Core Team, 2017). Scripts are available from OSF

(Bishop *et al.*, 2018b). One sample t-tests were used to investigate whether the LIs from each task were significantly different from zero, and matched two-sample t-tests assessed whether significant differences were found between tasks in order to test the hypothesis that individuals are consistent in measures of lateralization across different language domains. To confirm reliability of our measurements across all three tasks, Spearman's rank split-half correlations were calculated between odd and even trials. We also use Spearman correlations to explore the agreement between LIs for the three tasks. We had not planned any analyses of the effect of handedness, as we had only two left-handers in the sample, but we distinguish the left-handers visually in scatterplots.

Bland-Altman analysis (Altman & Bland, 1983) was used to assess the degree of agreement between the gold standard word generation task and the sentence generation task and, in doing so, test the hypothesis that laterality indices obtained from these two tasks were statistically equivalent. This method considers whether agreement between two measures is within the limits that might be expected on the basis of knowledge of reliability of the measures. In effect, it helps establish whether the correlation between two measures is compatible with the correlation of each measure with itself (i.e., its reliability). There is no *a priori* rationale for determining the largest limit of agreement that is acceptable (Giavarina, 2015), and so we based this on pre-existing data from previous studies by our group for comparing odd and even trials for the Word Generation Task in a large sample of children. For this sample, the mean difference between odd and even epochs was 0.16 with standard deviation of 1.24, which gives limits of agreement from -2.27 to 2.61. For the current study, we specified a limit of agreement from -2.5 to 2.5: if values on the Bland-Altman plot fall within these bounds, then the two tasks are deemed to be equivalent.

A second laterality index from the sentence generation task (SG2) was also computed using list generation as the comparison task. As noted above, there are various ways this could be done. The simplest approach is to compute the laterality index for each task separately, and then subtract the list generation LI from the sentence generation LI. Our pre-registered analysis specified we would conduct a one-tailed t-test with alpha level set to 0.025 to test the directional prediction that LI values from sentence generation with an active baseline will be larger than those with a resting baseline. Given that the LI is computed from the peak value, this may involve comparing LIs from different time windows. In additional exploratory analyses, we considered whether alternative methods might be preferable, to avoid this problem. One approach is to subtract the L-R difference wave obtained for the list generation task from the difference wave for the sentence generation task, and then compute a laterality index as usual, based on the peak value of this difference of differences waveform. A final approach, which might appear more compatible with fMRI, is to take the mean value of the difference wave within the period of interest for both SG and LG, and compute SG-LG. This approach does not require a peak to be identified.

In further exploratory analyses, we considered two questions: (a) whether the laterality index on a task was related to the

number of words produced, and (b) whether bimodality of laterality indices was an artefact of the method of computation, which involved identifying a peak. To do this, we compared the standard method of LI calculation with an alternative approach that involved subtracting the mean activation over the period of interest for the right channel from the mean for the left channel. We then compared LIs for the two approaches, to consider how far they agreed, and whether the latter method gave a more normal distribution of LIs.

Results
Descriptives
The summary of key results from each task is found in Table 1. Shapiro-Wilk tests confirmed that LI values from the three conditions were not normally distributed (LG: p = .024; SG: p = .030; WG: p = .004). Spearman's rho split-half correlations between odd and even trials from each task provide an index of the reliability of the LI values. The time-course of activation for each task is shown for each task in Figure 3. The distributions of LIs for each task are shown in Figure 4.

The data in Figure 4 are so striking that statistics are hardly necessary. Results of one-sample Wilcoxon tests are shown alongside the data, confirming that whereas the WG and SG tasks are strongly lateralised, the LG task is not. Lateralisation in the SG task was substantially higher than for the WG task.

Hypothesis 1: comparison of WG and SG tasks
All three tasks were significantly intercorrelated, as shown in Figure 5. Handedness is colour-coded, though we had too few left-handers to test for effects of handedness.

The mean LI values for the SG task were higher than those for the WG test (Mann-Whitney test, V = 471, p < .001). A Bland-Altman plot supported the conclusion that the tests were not interchangeable. As shown in Figure 6, the mean difference in LI values for the WG task versus the SG task (WG minus SG, dashed line) was outside of the predetermined limits of agreement (the pink shaded area). This is because LI values for SG tended to be higher than for WG.

Hypothesis 2: Subtracting list generation from sentence generation LI
Our second hypothesis concerned the question of whether we would see stronger lateralisation on the sentence generation task if we followed the fMRI method adopted by Mazoyer et al. (2014) and subtracted the activation from list generation. Because this kind of subtraction is not customary in fTCD studies, there is no agreed procedure, and accordingly, we explored three different approaches:

A) Compute a LI as usual for SG and LG, and subtract the LG value from the SG value. This involves using the LI values already derived for the analyses presented above, and is the method we specified in our preregistration.

B) Subtract the L-R difference waveform for LG from the L-R difference waveform for SG, and identify the peak difference, with the LI corresponding to the mean blood flow around the peak, as done previously.

C) Subtract the mean L-R flow in the period of interest for LG from the mean L-R flow in the same period for SG.

The pre-registered paired comparison between the original LI for SG and the LI with approach A was not statistically significant (Mann-Whitney test: V = 340, p = .073). Results from the other methods of computation are simply summarised: results obtained with the three different subtraction methods were highly intercorrelated (all correlations greater than .90), and in no case did the use of a subtraction procedure increase the LI for sentence generation. The correlations for the original LI for SG and the LIs using the three subtraction methods ranged from 0.66 to 0.79. Nevertheless, although the group mean and SD were similar for LIs with and without the subtraction, the subtraction had a clear impact on the rank ordering of LIs. Figure 7 shows LI data from the LG and SG tasks, plus SG with subtraction method A. In effect, with the subtraction method, those who have very similar LIs for SG and LG will see reduced LI-A, those with bilateral LI on LG will not change, whereas those with right-biased LI for LG and left-biased LI

Table 1. Descriptive statistics of functional transcranial Doppler sonography (fTCD) data. Mean (SD) statistics for each task, including number of trials, mean N words produced per trial, lateralization index (LI) and peak latency, the number of participants that were left, bilateral or right lateralised, and Spearman's rho for split-half (odd-even) reliability of the LI, with 95% bootstrapped CI.

	List	Sentence	Word
N trials	18.87 (1.36)	18.65 (1.20)	18.94 (1.12)
N words produced	10.12 (2.57)	9.13 (1.12)	4.34 (0.72)
LI	0.61 (2.04)	5.45 (2.21)	2.60 (1.55)
Latency	12.47 (2.77)	12.65 (2.76)	11.61 (2.86)
N left	10	30	23
N bilateral	17	0	7
N right	4	1	1
Split-half r	0.73 (.49-.87)	0.82 (.60-.94)	0.46 (.10-.70)
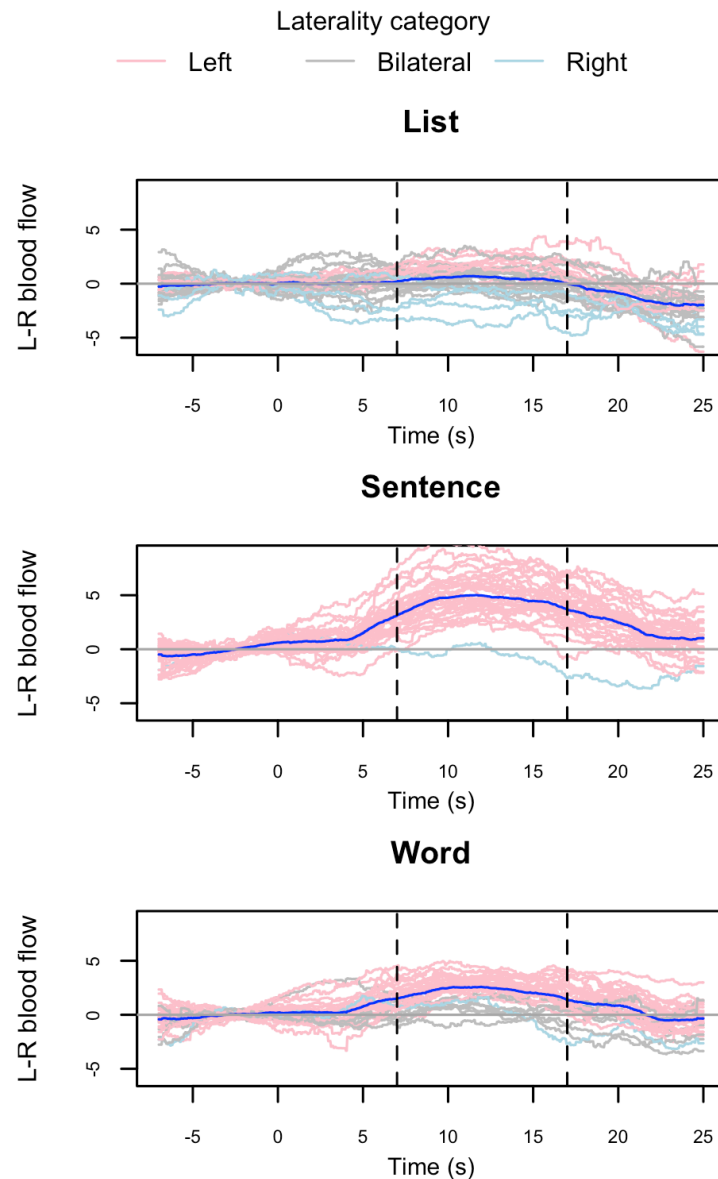
Figure 3. Time course of activation for each task. Bold line shows group mean, pale lines are individual participants, colour coded by laterality category for that task (red = left lateralised; grey = bilateral; blue = right lateralised).

for SG will become more left-biased on LI-A, because a negative number is subtracted from the raw LI.

Exploratory analysis
a) Does the strength of LI relate to the number of words produced in a task?
A question that is often raised is whether language laterality depends on how challenging a task is. We previously failed to show any association of amount of verbal material produced and LI on verbal tasks, when manipulating difficulty within-subjects (Badcock *et al.*, 2012b). In the current study, we observed higher LIs and more words produced in the Sentence Generation task than the Word Generation task, suggesting

that verbal production may influence strength of lateralisation (see Table 1). However, the number of words produced in the List Generation task was greater still, and this task had much weaker lateralisation. We also considered the Spearman correlations between N words produced and LI within each task. None of the correlations reached significance: LG, r_s = -0.19, p = .30; SG, r_s = 0.10, p = .61; WG, r_s = -0.06, p = .73. These results indicate that the differences in LI observed between tasks was not simply due to behavioural differences in the amount of speech produced. However, it is worth noting that the LI values were calculated from fTCD recordings during the speech generation (planning) phase, rather than overt speech production; in contrast, Gutierrez-Sigut *et al.* (2015) noted a
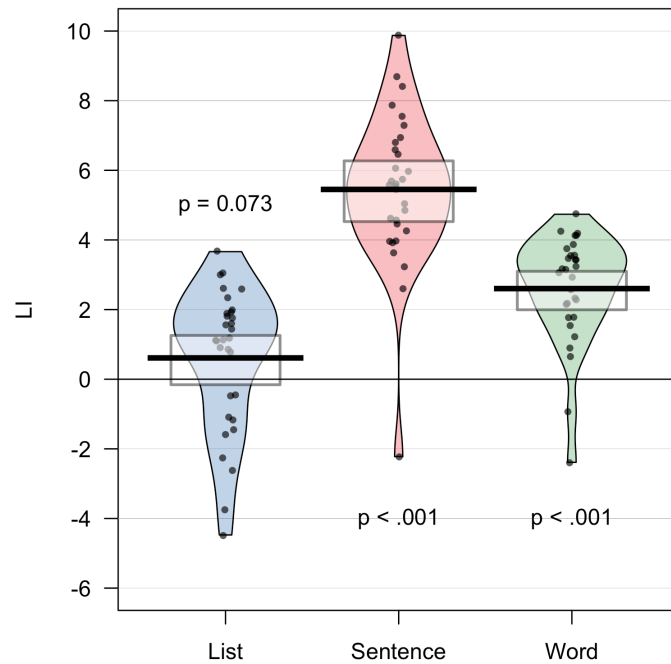
Figure 4. Pirate plot of lateralization index (LI) values. Pirate plot (Phillips, 2017) showing distributions of individual LI values for each task, with the bold line indicating the mean. The results of 1-sample Wilcoxon tests comparing the LI values to zero (no lateralisation) are also reported.
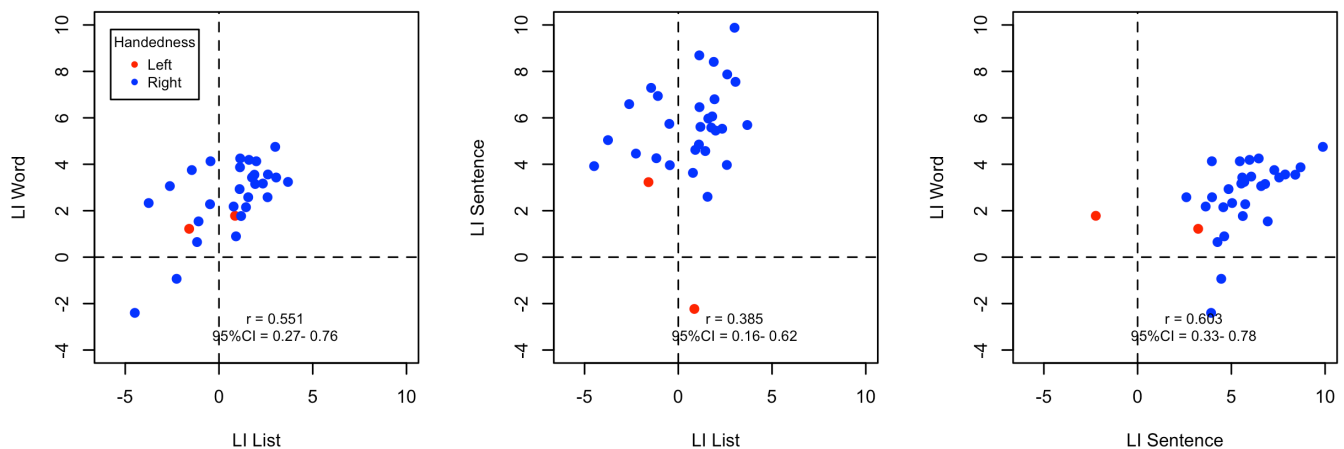
Figure 5. Pairwise scatterplots of lateralization index (LI) values. Scatterplots showing the pairwise comparisons of LI values between tasks. Spearman correlation coefficients (and 95% confidence intervals) are also reported. Right handed participants are indicated with blue dots; left handed participants with red dots.

difference between overt and covert word generation in word fluency tasks. For phonological fluency, there was a significant relationship between the amount of speech and LI values during an overt condition, but not when covert speech was used.

b) Is the bimodality of laterality indices an artefact of the method of calculation?

The standard method of computing a LI with fTCD was developed by Deppe and colleagues (2004), and involves finding

a peak in the difference wave and then computing the mean amplitude around that peak. This inevitably will induce bimodality in the distribution of LIs, as seen in Figure 4 for list generation, where a point of rarity occurs around zero. An alternative method of computing a laterality index was described above (method C) – taking the mean of the difference wave in the period of interest. Using this measure, LI values for the three conditions were as follows: for List Generation, LI = 0.48 (sd = 1.32); Sentence Generation LI = 4.58 (sd = 1.88); Word Generation, LI = 2.05 (sd = 1.17). Figure 8 shows the distribution
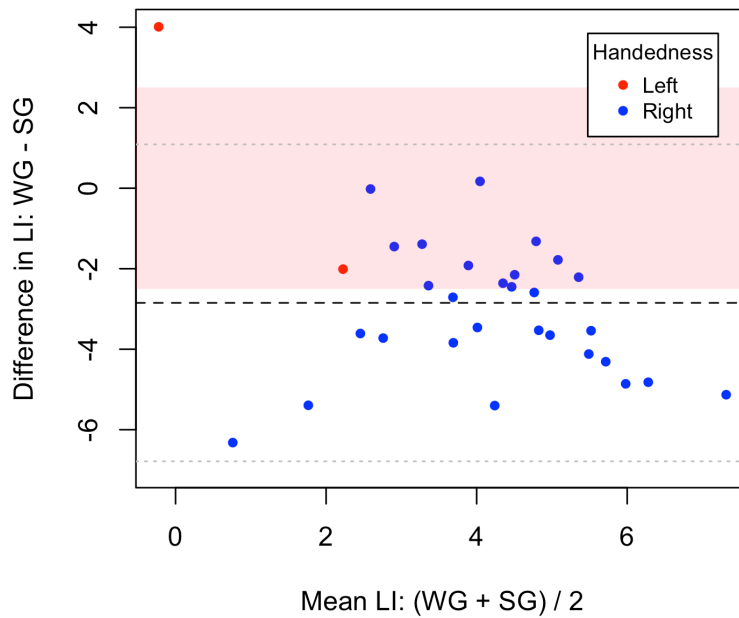
Figure 6. Bland-Altman plot comparing lateralization indices (LI's) for Sentence Generation and Word Generation. The pink shaded region denotes the predetermined limits of agreement. The dashed line is the mean difference between Sentence and Word Generation LI values, and the dotted lines are the 95% confidence interval around the mean. Right handed participants are indicated with blue dots; left handed participants with red dots.
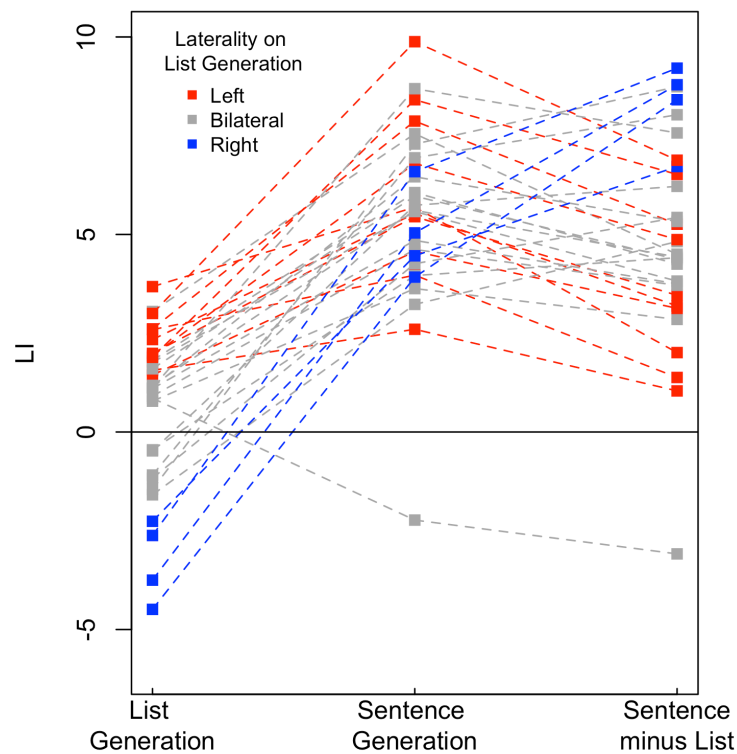
Figure 7. Effect of subtraction analysis on lateralization index (LI) values. Laterality indices for raw List Generation, raw Sentence Generation and the difference between Sentence Generation and List Generation for all individuals (using subtraction approach A).

LI from peak

LI from mean in POI
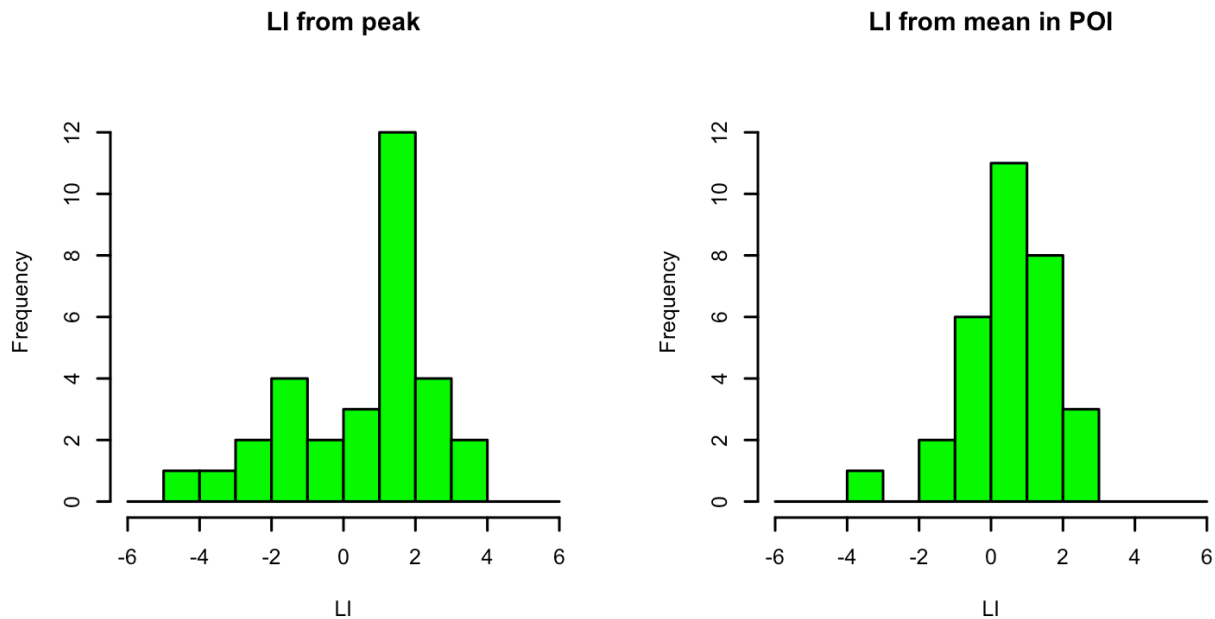
Figure 8. Impact of the method of calculating lateralization index (LI) on LI distribution. Distributions of LI values for list generation, computed from average around the peak, and mean over the period of interest.

of scores on the list generation task using this method compared with the traditional peak-based method. It is evident from inspection that the bimodality of the laterality index distribution is not seen when the means-based method is used. Despite the very different shapes of distributions from the two methods, they were highly intercorrelated: $r_s = 0.94$.

Discussion

The aims of this study were to determine (a) whether a laterality index from a sentence generation task was equivalent to the gold-standard word generation task, and (b) whether subtracting a laterality index from a comparison list generation task would enhance the laterality index. We compared the lateralisation indices of a group of 31 participants across three tasks: word generation (WG), sentence generation (SG) and list generation (LG). Regarding the first hypothesis, we found that the WG and SG tasks were not equivalent, as laterality was significantly stronger for the SG task. The correlation between the two tasks was moderately strong ($r_s = 0.60$). These findings lend some tentative support to the idea that language lateralisation is not a unitary function, and that in some individuals, different components of language may be preferentially processed in different hemispheres. Further support for this idea comes from a previous fMRI study comparing lateralisation as measured by WG and SG in a clinical population, which revealed that masses in the left-hemisphere Wernicke's area led to reduced lateralisation indices in this region for the sentence generation but not word generation, while the reverse was found for Broca's area (Partovi *et al.*, 2012). While this was a small study with task demands different to our own, it highlights a possible division of labour for the processing of different language components.

In fMRI studies, it is customary to subtract activation from a comparison task from activation associated with the language task of interest, with the aim of obtaining a purer measure that removes activation associated with more general motor or perceptual processing. To test our second hypothesis we considered whether we could achieve a similar effect by subtracting the LI from a list generation task from the sentence generation LI. The two tasks were well-matched in terms of the number of words spoken, but list generation did not show significant laterality at the group level. It was therefore not surprising that the subtraction had no effect on the laterality index for sentence generation at the group level.

An unanticipated finding, illustrated in Figure 5 and Figure 7, is that laterality on list generation was significantly correlated with laterality on the other two tasks. This was surprising: the overall lack of lateralisation at the group level might suggest that both hemispheres participate equally in this task, and any individual variation is not reliable or meaningful. The data, however, suggest that individuals do vary in the hemisphere used to generate over-learned lists, even though there is no bias to either side at the group level. The fact that the LIs from word and sentence generation tasks are correlated with list generation suggests that common processes of speech production are implicated in both tasks, with an additional left-sided bias superimposed that is specific to generation of novel, meaningful language.

This finding does not give a conclusive answer to the question of whether a list generation comparison task should be used in language laterality research: rather, it raises further questions. On the one hand, the logic of the comparison task subtraction

methodology used in fMRI in this area is supported: any task that involves language generation will implicate speech production mechanisms, which do not appear to be lateralised at the population level. If we subtract this activation, we should obtain a clearer index of lateralisation specific to semantic and syntactic processing. Note, however, that, as shown in Figure 7, in doing this, we will allocate more extreme lateralisation indices to those who are inconsistently lateralised for list and sentence generation: this is an inevitable consequence of the algebra of the subtraction. This suggests an alternative interpretation of the findings of Mazoyer *et al.* (2014) regarding individuals with extreme atypical right-sided language laterality: these are likely to be those who are left-sided for list generation but right-sided for sentence generation. It may be that inconsistency between lateralisation of speech and language is more important than the direction or size of laterality of either alone.

In an exploratory analysis, we investigated whether the conventional method of computing the LI from the peak difference between left and right CBFV may produce a bimodal distribution of LI values. Figure 8 shows that such a bimodal distribution was observed for the list generation task, where LI values tended to zero. The alternative method of computing LI from the mean left minus right difference in CBFV within the period of interest has the advantage of producing normally distributed values, without a point of rarity at LI=0. Although the choice of the most optimal method for calculating LI may vary from task to task, we suggest that the latter method is a more veridical representation of CBFV asymmetry when task performance is sustained throughout the period of interest.

Our study was focussed on characteristics of typical language lateralisation, and a consequent limitation was that we had only two left-handed participants. The fact thatone of these left-handers was a clear outlier on the SG task may be of interest for future investigation. To explore this further, future work might investigate the relationship between tasks in a sample with more left-handers, using more quantitative assessments of handedness and language ability.

The technological limitations of the fTCD technique also cannot be overlooked – as mentioned above, poor spatial resolution means that functions can be localised only at a hemispheric level and the requirement of a suitable temporal window means a satisfactory signal cannot be obtained in around 5% of participants (Lohmann *et al.*, 2006). Indeed, we had two such exclusions whereby suitable window could not be found, and one individual from which data could not be obtained as set-up was aborted.

Conclusion
We adapted the sentence generation task of Mazoyer *et al.* (2014) for use with fTCD and demonstrated that it gives robust left-lateralisation in a group of unselected individuals, with good split-half reliability. The lateralisation on this task was significantly stronger than that seen on the gold-standard word generation task, although LIs from the two tasks were correlated. An unanticipated finding was that the LI from a list generation task, which was used as a comparison task for sentence generation, was significantly correlated with the LIs from both word generation and sentence generation, even though at the group level, this task was not lateralised. Subtraction of the list generation LI from sentence generation LI had no impact on the mean sentence generation LI, though it did affect the rank ordering of LI values. Those with extreme values on the subtracted Sentence-List LI will be those who have discrepant laterality across the two tasks.

In sum, we conclude that the sentence generation task is a reliable and feasible task for use with fTCD, which may be useful for identifying individuals who are inconsistent in lateralisation across different language domains. There may be value also in further studies with list generation, which allows one to study laterality of speech production when there is no role for using semantic or syntactic processes to generate novel sequences. The lack of an overall lateral bias on list generation might suggest that both hemispheres participate equally in speech production, but our data indicate that there are meaningful individual differences in left- or right-sided bias on this task, whose functional significance remains to be established.

Data availability
All raw data, analysis scripts and processed data can be found on Open Science Framework.

Data:
CANDICE A2a: Measurement of language laterality, https://doi.org/10.17605/OSF.IO/68KEZ (Bishop *et al.*, 2018a)

This dataset is available under a CC0 1.0 Universal license.

Task scripts:
CANDICE A2a: Measurement of language laterality, https://doi.org/10.17605/OSF.IO/EYVDB (Woodhead *et al.*, 2018)

This dataset is available under a CC0 1.0 Universal license.

R analysis scripts:
CANDICE A2a: Measurement of language laterality, https://doi.org/10.17605/OSF.IO/X5DWZ (Bishop *et al.*, 2018b)

This dataset is available under a CC0 1.0 Universal license.

References

Altman DG, Bland JM: **Measurement in medicine: the analysis of method comparison studies.** *The Statistician.* 1983; **32**(3): 307–317.
Publisher Full Text

Badcock NA, Holt G, Holden A, *et al.*: **dopOSCCI: a functional transcranial Doppler ultrasonography summary suite for the assessment of cerebral lateralization of cognitive function.** *J Neurosci Methods.* 2012a; **204**(2): 383–388.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Badcock NA, Nye A, Bishop DV: **Using functional transcranial Doppler ultrasonography to assess language lateralisation: Influence of task and difficulty level.** *Laterality.* 2012b; **17**(6): 694–710.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Binder JR, Swanson SJ, Hammeke TA, *et al.*: **A comparison of five fMRI protocols for mapping speech comprehension systems.** *Epilepsia.* 2008; **49**(12): 1980–1997.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Bishop D, Thompson PA, Woodhead Z, *et al.*: **Data.** *Open Science Framework.* 2018a.
http://www.doi.org/10.17605/OSF.IO/68KEZ

Bishop D, Thompson PA, Woodhead Z, *et al.*: **Scripts.** *Open Science Framework.* 2018b.
http://www.doi.org/10.17605/OSF.IO/X5DWZ

Bradshaw AR, Bishop DVM, Woodhead ZVJ: **Methodological considerations in assessment of language lateralisation with fMRI: a systematic review.** *PeerJ.* 2017; **5**(7): e3557.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Brainard DH: **The Psychophysics Toolbox.** *Spat Vis.* 1997; **10**(4): 433–436.
PubMed Abstract | **Publisher Full Text**

Bruckert L: **Is language laterality related to language abilities?** 2016.
Reference Source

Deppe M, Knecht S, Papke K, *et al.*: **Assessment of hemispheric language lateralization: a comparison between fMRI and fTCD.** *J Cereb Blood Flow Metab.* 2000; **20**(2): 263–268.
PubMed Abstract | **Publisher Full Text**

Deppe M, Ringelstein EB, Knecht S: **The investigation of functional brain lateralization by transcranial Doppler sonography.** *Neuroimage.* 2004; **21**(3): 1124–1146.
PubMed Abstract | **Publisher Full Text**

Dräger B, Knecht S: **When finding words becomes difficult: is there activation of the subdominant hemisphere?** *Neuroimage.* 2002; **16**(3 Pt 1): 794–800.
PubMed Abstract | **Publisher Full Text**

Gagnon L, Sakadžić S, Lesage F, *et al.*: **Quantifying the microvascular origin of BOLD-fMRI from first principles with two-photon microscopy and an oxygen-sensitive nanoprobe.** *J Neurosci.* 2015; **35**(8): 3663–75.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Gaillard WD, Balsamo L, Xu B, *et al.*: **fMRI language task panel improves determination of language dominance.** *Neurology.* 2004; **63**(8): 1403–1408.
PubMed Abstract | **Publisher Full Text**

Giavarina D: **Understanding Bland Altman analysis.** *Biochem Med (Zagreb).* 2015; **25**(2): 141–151.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Grabitz CR, Watkins KE, Bishop DVM: **Cerebral lateralisation of first and second languages in bilinguals assessed using functional transcranial Doppler ultrasound [version 1; referees: 1 approved, 1 approved with reservations, 1 not approved].** *Wellcome Open Res.* 2016; **1**: 15.
Publisher Full Text

Gutierrez-Sigut E, Payne H, MacSweeney M: **Investigating language lateralization during phonological and semantic fluency tasks using functional transcranial Doppler sonography.** *Laterality.* 2015; **20**(1): 49–68.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Illingworth S, Bishop DV: **Atypical cerebral lateralisation in adults with compensated developmental dyslexia demonstrated using functional transcranial Doppler ultrasound.** *Brain Lang.* 2009; **111**(1): 61–65.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Kleiner M, Brainard D, Pelli D: **What's new in Psychtoolbox-3?** *Perception.* 2007; **36**: ECVP Abstract Supplement.
Reference Source

Knecht S, Deppe M, Ebner A, *et al.*: **Noninvasive determination of language lateralization by functional transcranial Doppler sonography: a comparison with the Wada test.** *Stroke.* 1998a; **29**(1): 82–86.
PubMed Abstract | **Publisher Full Text**

Knecht S, Deppe M, Ringelstein EB, *et al.*: **Reproducibility of functional transcranial Doppler sonography in determining hemispheric language lateralization.** *Stroke.* 1998b; **29**(6): 1155–1159.
PubMed Abstract | **Publisher Full Text**

Krach S, Hartje W: **Comparison of hemispheric activation during mental word and rhyme generation using transcranial Doppler sonography.** *Brain Lang.* 2006; **96**(3): 269–279.
PubMed Abstract | **Publisher Full Text**

Lohmann H, Ringelstein EB, Knecht S: **Functional transcranial Doppler sonography.** *Baumgartner RW (ed): Handbook on Neurovascular Ultrasound, Front Neurol Neurosci.* 2006; **21**: 251–260.
PubMed Abstract | **Publisher Full Text**

Lust JM, Geuze RH, Groothuis AG, *et al.*: **Driving performance during word generation--testing the function of human brain lateralization using fTCD in an ecologically relevant context.** *Neuropsychologia.* 2011; **49**(9): 2375–2383.
PubMed Abstract | **Publisher Full Text**

MATLAB and Statistics Toolbox Release. The MathWorks, Inc., Natick, Massachusetts, United States. 2012b.

Mazoyer B, Zago L, Jobard G, *et al.*: **Gaussian mixture modeling of hemispheric lateralization for language in a large sample of healthy individuals balanced for handedness.** *PLoS One.* 2014; **9**(6): e101165.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Partovi S, Jacobi B, Rapps N, *et al.*: **Clinical standardized fMRI reveals altered language lateralization in patients with brain tumor.** *AJNR Am J Neuroradiol.* 2012; **33**(11): 2151–2157.
PubMed Abstract | **Publisher Full Text**

Payne SJ: **Cerebral Blood Flow and Metabolism.** Singapore: Word Scientific Publishing Co. Pte. Ltd. 2018.
Publisher Full Text

Phillips N: **A Companion to the e-Book "YaRrr!: The Pirate's Guide to R"** (R package version 0.1.5). 2017.
Reference Source

R Core Team: **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. 2017.
Reference Source

Somers M, Neggers SF, Diederen KM, *et al.*: **The Measurement of Language Lateralization with Functional Transcranial Doppler and Functional MRI: A Critical Evaluation.** *Front Hum Neurosci.* 2011; **5**: 31.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Stroobant N, Buijs D, Vingerhoets G: **Variation in brain lateralization during various language tasks: a functional transcranial Doppler study.** *Behav Brain Res.* 2009; **199**(2): 190–196.
PubMed Abstract | **Publisher Full Text**

Tailby C, Abbot DF, Jackson GD: **The diminishing dominance of the dominant hemisphere: Language fMRI in focal epilepsy.** *NeuroImage Clin.* 2017; **14**: 141–150.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wada J, Rasmussen T: **Intracarotid injection of sodium amytal for the lateralization of cerebral speech dominance.** *J Neurosurg.* 1960; **17**(2): 266–282.
Publisher Full Text

Whitehouse AJ, Bishop DV: **Cerebral dominance for language function in adults with specific language impairment or autism.** *Brain.* 2008; **131**(Pt 12): 3193–3200.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wilson AC, Bishop DVM: **Resounding failure to replicate links between developmental language disorder and cerebral lateralisation.** *PeerJ.* 2018; **6**: e4217.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Woodhead ZVJ, Bishop DVM, Thompson PA, *et al.*: **Task.** *Open Science Framework.* 2018.
http://www.doi.org/10.17605/OSF.IO/EYVDB

Open Peer Review

Current Peer Review Status: ✓ ✓ ✓ ?

Version 3

Reviewer Report 06 May 2020

https://doi.org/10.21956/wellcomeopenres.17247.r38111

? **Andreas Jansen** (iD)

Department of Psychiatry and Psychotherapy, Philipps University of Marburg, Marburg, Germany

Woodhead *et al*. assessed language lateralization using functional Doppler sonography (fCTD). Language lateralization was described by a laterality index (LI). They investigated the impact of different language tasks as well as the impact of different calculation methods on the LI.

This is an interesting study, assessing important methodological aspect of fCTD. The study has several strengths, among others the preregistered analyses and the disclosure of all analysis methods. It is quite interesting to see that the Deppe *et al.* (1997) approach (summarized in Deppe *et al.* (2004)) tends to overestimate (the absolute value) of the LI.

In my opinion however, some aspects of the study are poorly described. I will try to explain my impression in the following.

The **abstract** is poorly written. Many statements are too unspecific. It is difficult for the reader to understand the aim, the results and the conclusions based on the information from the abstract alone. To give some examples:
Example 1: "Language laterality has mostly been studied with fTCD using a word generation task, but it is not clear whether this is optimal…" and this was a pre-registered methodological study designed to explore novel approaches to optimising measurement of language lateralisation using fTCD…". What does "optimal" mean in this context? A task that is as short as possible? A task that yields more reliable LIs than previous tasks? A task that yields LIs that are more strongly lateralized? A task has a higher agreement with other technique, e.g. fMRI or the Wada test?
Example 2: "Bland-Altman analysis showed that these two methods were not equivalent." It is impossible to understand what that means when reading only the abstract. The authors might want to describe their hypotheses in more detail also in the abstract.
Example 3: "The comparison list generation task was not systematically lateralised, but nevertheless laterality indices (LIs) from this task were significantly correlated with the other two tasks." Again, when reading only the abstract the reader does not get useful information.

Since I am myself not a native English speaker I will refrain from any specific advice how to improve the

abstract. I suggest however that the authors rewrite the abstract so that is is understandable on its own.

I also suggest that the authors rewrite some parts of the **background**. First, some paragraphs make statements that are rather vague. Second, some information given are too uncritical with regard to the properties of fTCD.

To give <u>one example for unspecific statements</u>: "It is well-established that language is predominantly processed in the left cerebral hemisphere in the majority of people. The functional significance of this bias is not understood." What is meant by the second sentence? It is a statement that is so unspecific that it is almost meaningless.

To give <u>some examples for an uncritical description of the fCTD literature</u>: The authors state that "the word generation task (Knecht *et al*., 1998a; Knecht *et al*., 1998b) has been used as the gold standard method for assessing cerebral lateralisation using fTCD – with good concordance with fMRI (Deppe *et al*., 2000; Somers *et al*., 2011) and Wada technique (Knecht *et al*., 1998a; Wada & Rasmussen, 1960), and adequate test-retest reliability (Knecht *et al*., 1998a)."
The literature cited, in particular the early fTCD studies, have several methodological shortcomings. To give two examples from the studies that are mentioned:
Deppe *et al.* (2000) calculated LIs for a word generation task using both fMRI and fTCD. Aim of the authors was to assess the concordance of both methods, not only with regard to the hemispheric dominance (left/right/bilateral), but also with regard to the degree of lateralization. They reported a correlation of r=0.95 between both techniques. If you take a closer look at this publication, you will see that there remain open questions. How was the LI for fMRI data calculated? What statistical threshold was used? Why is the correlation between both techniques so high although the authors chose a calculation method for the LI-fMRI that has been shown to be not reliable (with regard to the degree of lateralization) when subjects are measured twice with fMRI? In my opinion, the results of this study do not fit together with many other studies, so its validity must be doubted. If the authors disagree, they might want to explain why the correlation between fTCD and fMRT is so much higher than for repeated fMRT measurements.

Knecht *et al.* (1998) compared fTCD with the Wada test. They concluded that their study "validates fTCD as a noninvasive and practical tool for the determination of language lateralization that can be applied for clinical and investigative purposes." When you want to apply fTCD for clinical purposes however you have to be able to assess whether a subject shows typical or atypical language dominance. Atypical language dominance includes both bilateral as well as right-hemispheric dominance. Why did Knecht and colleagues not include this category in their analysis? Is the correlation reported (r=0.92) also that high when only including subjects with typical language lateralization?

I cannot go into detail within the scope of this review. But there are many reasons why in particular the early fTCD publications overestimate the reliability of the technique and its concordance with other imaging modalities dramatically. I therefore think that statements such as "good concordance with fMRI" or "adequate test-retest reliability" should be used with caution.

The – preregistered - **hypotheses** that are stated at the end of the introduction are with regard to their verbal formulation is not sufficiently specific. What is for instance meant by "individuals will be consistent in measures of hemispheric lateralization across different language domains". Do you expect that if a subject is left-dominant for one task she is also left-dominant for another task? Or do you expect that a subject that is more strongly left-lateralized than another subject as assessed by task 1 is also more strongly left-lateralized when using task 2? These are different scenarios!

To give another example: Do you intend to test hypothesis 2 by comparing the mean LI (averaged across all subjects) between both implementations or on a single subject level? The authors might try to specify their hypotheses more clearly on a verbal level. Alternatively, they should describe more clearly in the methods section which analysis is used to test which hypothesis. At present, it takes some time for the reader to clearly disentangle which analysis is used for which reason. To me, it did not become fully clear. In my personal opinion, the manuscript would benefit from a more structured presentation of the analyses.

The **power analysis** calculated the required sample size to detect lateralisation greater than zero. Can it be that this power analysis is no way related to the specific questions stated in the introduction?

Descriptives: Do the authors expect that the LIs are normally distributed? Typically, you expect a bimodal distribution (see e.g. Knecht 2000), some are left-dominant, some are right-dominant. When considered separately, the LIs for subjects with left- and right-dominance might be normally distributed. Does it make sense to test for normality for the total sample?

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cognitive neuroscience, functional neuroimaging.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 16 March 2020

✔ **Guy Vingerhoets** iD

Department of Experimental Psychology, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium

The study aims to contribute to two important questions in laterality research: is sentence generation equivalent to word generation when it comes to the determination of hemispheric asymmetry of language laterality and is task subtraction used in fMRI also applicable in fTCD?

This clearly written paper has been commented on by two reviewers already, and their suggestions have been included in adapted versions of the paper. In addition, a problem with the original analysis has been corrected. The amendment resulted in a significantly stronger laterality index for the sentence generation task than for the word generation task.

Given all this previous work fine-tuning the paper, it appears to me that the available data have been analyzed thoroughly and that the conclusions are supported by the results obtained.
So, task subtraction does not enhance the laterality index in fTCD and sentence generation produces stronger laterality indices than word generation.
The former finding does not lead to a specific recommendation though, as subtraction might have advantages as well as disadvantages.

With regard to the latter finding, I was wondering whether this finding might influence the authors' opinion with regard to the status of word generation as the gold-standard for language laterality.

In the abstract the authors mention that it is unclear whether the use of a word generation task in fTCD is optimal for measuring language laterality. What is meant by optimal? If a stronger laterality index is better, should we then change the gold-standard to sentence generation? Or should we only acknowledge that different language tasks have different LI's but agree to keep on using word generation as the prototypical language task giving its widespread use and validation?

In sum, I believe this manuscript provides valuable information that contributes to our knowledge on the measurement of language laterality and can be approved for indexing.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Neuropsychology, Brain asymmetry, Laterality, Hemispheric functional segregation, Brain organization.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 03 October 2018

https://doi.org/10.21956/wellcomeopenres.16036.r33764

✔ **Eva Gutierrez-Sigut** 🆔
   University of Valencia , Valencia, Spain

This is a clearly written and needed study that aims to optimise the measurement of Language lateralisation using fTCD.

The study assess lateralisation indices of 31 participants during 3 different covert language production tasks, including the gold standard task (word generation). Results show that although similar and highly correlated, word generation and generation of sentences from pictures don't seem to be interchangeable tasks. More interestingly, the pattern of results that arises when "list repetition" is used as comparison task suggests that word and sentence generation add a directional bias to a continuum of laterality. Intraindividual consistency in lateralisation between different different language tasks might be explanatory of how language processing occurs in the brain,

I think that the authors efforts to constantly improve the pipeline of data pre-processing is commendable. Specifically the addition of an extra interpolation step which has the potential to result in lower number of valid trials rejected. I am curious of why the interpolations is performed after and not before down-sampling (which should theoretically provide better interpolation).

The statistical analyses presented, the pre-registered and the exploratory, are clearly motivated and advance the field. Can the authors offer any recommendation on best practice for future studies based on these comparisons? e.g. is it better to avoid the bimodality in LI distributions?

Although I can understand where hypothesis 2 comes from it could be more motivated in the intro. On the same note, the focus on the Mazoyer et al. (2014)[1] results is put on how the strongly right lateralised individuals were left handed. I am unsure of how this ties up with the present study. The reasoning might need a bit of unpacking.

It is stated that the word and sentence generation tasks are not interchangeable, this is based on a very small percentage of participants (6 including the only two left handed participants) falling on the non-acceptable range for the Bland-Altman plot. Are the results from the two tasks really that different? If they are, what do the authors think that might be the difference? grammatical/syntactic processing that tends to be more frontal? task engagement? I think the authors propose a really interesting framework for future fTCD research and further discussion on what particular aspects of the tasks could affect LIs would be really valuable.

The correlations between behaviour and LI are not significant, however this is not strange taking into account that the behaviour inputed to the correlation is for the report phase and LI is being measured before.

Very minor point:

Although the authors say the they have substituted the term "baseline task" by "comparison task" to avoid confusion I think baseline is used to refer to task further down the manuscript (see below). This might be intentional.

"...LI values from sentence generation with an active baseline will be larger than those with a resting baseline..."

"...This finding does not give a conclusive answer to the question of whether a list generation baseline should be used in language laterality research: rather, it raises further questions...."

References

1. Mazoyer B, Zago L, Jobard G, Crivello F, et al.: Gaussian mixture modeling of hemispheric lateralization for language in a large sample of healthy individuals balanced for handedness.*PLoS One*. 2014; **9** (6): e101165 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cognitive neuroscience; Deafness; Lateralisation; ERPs; Reading; Reading
Difficulties

**I confirm that I have read this submission and believe that I have an appropriate level of
expertise to confirm that it is of an acceptable scientific standard.**

Author Response 11 Oct 2018
Zoe Woodhead, University of Oxford, Oxford, UK

We would like to thank the reviewer for their thoughtful and constructive comments, which we have
addressed in the revised version of the manuscript as discussed below.

1. Interpolation of fTCD artefacts
We observe that spike artefacts in our fTCD are very brief – they may only last one or two samples.
This means that there are, inevitably, fewer artefacts in the downsampled data than the raw data.
Hence, it makes sense to perform the interpolation of spike artefacts after downsampling.

2. Recommendations for use of LIs in future research
As we report in the paper, using the mean L-R difference within the POI as a measure of LI (the
'mean-based' LI) results in more normally distributed LI values, and avoid the bias away from LI=0,
which is a problem with conventional peak-based LIs. If your subsequent analyses rely on the
assumption of normality, then mean-based LIs are definitely a better option. We also believe that it
is a better reflection of the actual lateralisation of brain activity, as it is calculated from CBFV
throughout the period of interest. In some individuals, we have observed that peak-based LI values
can be driven by a transient period of lateralised CBFV that is not a real reflection of their overall
pattern of CBFV across the epoch.
The decision of which measure to use may depend on the task in question – some tasks that
require a transient event that occurs at a variable timepoint within the period of interest may be
better suited to a peak-based method – but in general, we have started to use a mean-based
measure as standard in our research.
We have updated the discussion to make this recommendation clearer to the reader.

3. Motivation for hypothesis
We have reorganised the Background section to highlight the key points from Mazoyer's paper
(Mazoyer et al., 2014), and have deleted some extraneous detail about this study. We hope that
this makes the rationale behind Hypothesis 2 easier to understand.

4. Interchangeability of word and sentence generation tasks
We agree with the reviewer that it is hard to interpret the results from the Bland-Altman plot: the
data do not support a conclusive interpretation. The LI values from the two tasks are not
significantly different, but equally the Bland-Altman task (which was defined in our preregistration
as the method we would use to assess similarity) identified some participants for whom the LI

values differed between the two tasks. The data suggest that for most people, word and sentence generation have very similar levels of lateralisation, but that this isn't the case for everyone. This may reflect meaningful heterogeneity in the population of typically developed adults in the organisation of language functions across the two hemispheres. The current data do not allow us to speculate further on this matter, but a more recent study we have conducted using a range of six different language tasks (Woodhead et al, 2018) suggests that this heterogeneity does exist, and (in line with Mazoyer's findings) that an atypical pattern of language lateralisation can occur in a minority of participants.

5. Correlations with task performance
We agree the amount of speech produced is less likely to be correlated with LI when it is recorded during covert speech than when overt speech is used (as was observed in Gutierrez-Sigut et al., 2015). Our main purpose in reporting these statistics is because there was a large difference between conditions in the amount of speech produced. We therefore wished to note that this behavioural difference was unlikely to have driven the results we saw between tasks in LI values. We have clarified this motivation in the Results section. Inclusion of correlations within individuals was just included for completeness: we suspected that if we did not report this, readers may have been curious about it. We agree it is not surprising to find no significant associations.

6. Use of 'baseline task' or 'comparison task'
Thank you for pointing out this inconsistency! We have rectified the error and made sure that 'comparison task' is now used throughout.

References
Gutierrez-Sigut, E., Payne, H., & MacSweeney, M., Investigating language lateralization during phonological and semantic fluency tasks using functional transcranial Doppler sonography, *Laterality: Asymmetries of Brain, Body and Cognition*, **20**, 49-68 (2015).

Mazoyer B., Zago, L., Jobard, G., Crivello, F., Joloit, M.m Perchey, G., Mellet, E., Petit, L., Tzourio-Mazoyer, N., Gaussian Mixture Modeling of hemispheric lateralization for language in a large sample of healthy individuals balanced for handedness. *PLOS One*, **9** (2014).

Woodhead ZVJ, Bradshaw AR, Wilson AC, Thompson PA, Bishop DVM., Testing the unitary theory of language lateralisation using functional transcranial Doppler sonography in adults. *bioRxiv,* **437939,** doi: https://doi.org/10.1101/437939, (2018).

Competing Interests: No competing interests were disclosed.

Reviewer Report 21 September 2018

https://doi.org/10.21956/wellcomeopenres.16036.r33849

Ervin Sejdić (iD)

Department of Electrical and Computer Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA, USA

- A different statistical test may be needed, unless the authors have confirmed that LI points adhere to assumptions of t-tests.
- Is it possible also to include mean values for observed blood flow velocities?
- While LI is a typical feature to understand lateralisation, have you considered other signal features? LI is an intuitive measure, but also a very crude measure.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 11 Oct 2018
Zoe Woodhead, University of Oxford, Oxford, UK

In response to the first point, we have reported Shapiro-Wilk normality tests for the LI data, which did not indicate that LI values from any of the three tasks were non-normally distributed. This may seem odd, given the bimodal distribution observed for the List Generation task in Figure 4. The Shapiro-Wilk test result for this condition was p=.084. Statistically, it is impossible to say for certain whether the data is truly normally distributed, or whether low power (due to the relatively small sample size) affected the Shapiro-Wilk result. But intuitively, the method of identifying LI values from the peak in L-R difference seems bound to result in a bimodal distribution if the 'true' mean lateralisation is close to zero; hence, we feel justified in recommending a change in the way the LI values are calculated, despite the apparent normality of the conventional LI values reported here.

For the second point, we have now reported the mean LI values using the alternative 'means-based' method at the end of the Results section. They were as follows: for list generation,

LI = 0.16 (sd = 1.83); sentence generation LI = 2.48 (sd = 1.70); word generation, LI = 1.91 (sd = 1.30).

For the final point, we agree with the reviewer that there may be other features of interest in the fTCD signal that are not part of the conventional analysis methodology; for example, examining the timecourse of blood flow may be a fruitful avenue of investigation (e.g. Meyer et al., 2014). However, our aim in the current study was to examine the validity of current standard practise in fTCD studies, and the analyses we carried out to achieve this aim were planned in our preregistration. We would like to conduct further research on whether more novel methods may give additional insights into lateralisation, as the reviewer suggests.

Reference
Meyer G., Spray A., Fairlie J., Uomini N., Inferring common cognitive mechanisms from brain blood-flow lateralization data: A new methodology for fTCD analysis. *Frontiers in Psychology,* **5**, 1-15. (2014).

Competing Interests: No competing interests were disclosed.