Cerebral lateralisation of first and second languages in bilinguals assessed using functional transcranial Doppler ultrasound [version 1; peer review: 1 approved, 1 approved with reservations, 1 not approved]

Clara R. Grabitz, Kate E. Watkins, Dorothy V. M. Bishop

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

Abstract

Background: Lateralised representation of language in monolinguals is a well-established finding, but the situation is much less clear when there is more than one language. Studies to date have identified a number of factors that might influence the brain organisation of language in bilinguals. These include proficiency, age of acquisition and exposure to the second language. The question as to whether the cerebral lateralisation of first and second languages are the same or different is as yet unresolved.

Methods: We used functional transcranial Doppler sonography (FTCD) to measure cerebral lateralisation in the first and second languages in 26 high proficiency bilinguals with German or French as their first language (L1) and English as their second language (L2). FTCD was used to measure task-dependent blood flow velocity changes in the left and right middle cerebral arteries during word generation cued by single letters. Language history measures and handedness were assessed through self-report questionnaires.

Results: The majority of participants were significantly left lateralised for both L1 and L2, with no significant difference in the size of asymmetry indices between L1 and L2. Asymmetry indices for L1 and L2 were not related to language history, such as proficiency of the L2.

Conclusion: In highly proficient bilinguals, there is strong concordance for cerebral lateralisation of first and second languages.

Keywords
Laterality, Bilingualism, FTCD
Introduction
The two cerebral hemispheres of the brain are neither structurally nor functionally identical. Hemispheric specialisation reflects a variety of factors influencing brain development, including genetics, development, experience and pathology. Language ability is particularly interesting in this regard, since, at least in monolinguals, it is predominantly left lateralised (Knecht et al., 1998a). The representation of language in the bilingual brain has been a topic of controversies. While cases of differential recovery patterns for individual languages in stroke patients point towards separate neural representations, neuroimaging of healthy individuals has mostly reported the involvement of overlapping cortical areas for both first and second languages (Abutalebi et al., 2005; Paradis, 2004).

The neural organisation of language in multilinguals is affected by age of acquisition, proficiency and exposure effects (for neuroimaging and behavioural reviews see Abutalebi et al., 2005; Hull & Vaid, 2007; Perani & Abutalebi, 2005). Typically, differential activation has been reported for late acquisition or low proficiency groups, though the impact of these individual differences seems to be task dependent (Dehaene & Cohen, 1997; Kim et al., 1997; Klein et al., 1995; Klein, 2003; Wartenburger et al., 2003).

A range of methods has been used to assess anatomical and functional differences between cerebral hemispheres, and depending on experimental aims as well as environmental and task constraints, some methods may be more favourable than others. Here, our focus was on comparing cerebral lateralisation for first and second languages using functional transcranial Doppler ultrasonography (FTCD). This method uses ultrasound to measure cerebral blood flow velocity (CBFV) in the left and right hemispheres. The change in CBFV reflects the task dependent contribution of each hemisphere due to neurometabolic coupling i.e. brain areas showing task-dependent neuronal firing need to replenish metabolic resources, requiring increased blood flow (Aaslid et al., 1982; Deppe et al., 2004). In order to assess language lateralisation, the middle cerebral artery (MCA), which supplies extensive regions of the cortex, including frontal, temporal and parietal areas, is insonated (van der Zwan et al., 1993). These cortical regions in the left hemisphere contain areas that are necessary for language processing and production, including Broca’s and Wernicke’s areas in the inferior frontal and superior temporal lobes, respectively. FTCD is a reliable and valid measure of language lateralisation, giving good correlations with the gold standard intracarotid amobarbital test and functional MRI (fMRI) (Bishop et al., 2009; Deppe et al., 2004; Groen et al., 2012; Illingworth & Bishop, 2009; Knake et al., 2003; Knecht et al., 1998a; Knecht et al., 1998b; Rihs et al., 1999; Somers et al., 2011; Stroobant et al., 2011).

Compared to other methods used to determine lateralisation, FTCD offers a variety of advantages; it is inexpensive, non-invasive, comfortable, easily applicable, mobile and child-friendly (Bishop et al., 2010; Knecht et al., 1998b). This makes it useful for repeated assessment of language lateralisation. While FTCD has been used to study cerebral lateralisation in monolinguals, it has not, to our knowledge, been used to compare lateralisation of two languages in bilingual participants, i.e. people who use more than one language on a regular basis (Grosjean, 1989).

In the current experiment, we used the cued word generation task, which is a well validated and commonly used productive language task (Knecht et al., 1998a; Knecht et al., 1998b), to test whether there are differences in language lateralisation between first (L1) and second (L2) languages in highly proficient bilinguals. A secondary aim was to establish whether lateralisation of L2 was influenced by language history variables, such as age of acquisition. We predicted that the extent of left lateralisation of bilingual speakers would relate to their proficiency levels. In addition, our bilingual participants were native speakers of either French or German, making it possible to see whether there was any indication of language-specific effects, as might be expected if language representation were affected by a similarity between L1 and L2. For instance, on a measure of lexical similarity, English has 60% similarity with German, but only 27% with French (https://www.ethnologue.com/language/eng).

Method
Participants
Participants were recruited through the Oxford University German Society and Oxford University French Society, as well as through posters in the Experimental Psychology building. Participants were aged over 18 years and were either German-English or French-English bilinguals. All had normal or corrected to normal vision. Individuals with a diagnosis of any speech, language or learning impairment, affected by a neurological disorder or taking medication affecting brain function e.g. antidepressants, were not included in the study. A total of 40 individuals were assessed for viability as study participants. In total, 14 participants were excluded for a range of reasons, including no suitable Doppler signal, due to the inability to find a suitable temporal window in the skull, or failure to stabilize the Doppler signal for the required amount of time (11 participants), or low quality data (3 participants). The final sample consisted of 26 individuals (M age = 22.81 years, SD = 3.63; 19 women, 7 men), of whom 11 reported French as their L1 and 15 reported German as their L1.

Ethics statement
The study was approved by the University of Oxford Central Research Ethics Committee (CUREC), approval number, MS-IDREC-C1-2015-126). All participants provided written informed consent.

Apparatus
A commercially available transcranial Doppler ultrasonography device (DWL, Multidop T2; manufacturer, DWL Elektronische Systeme, Singen, Germany) was used for continuous measurements of the changes in CBFV through the left and right MCA. The MCA was insonated at ~5 cm (40–60 mm). Activity in frontal and medial cortical areas, supplied by the anterior cerebral artery, and inferior temporal cortex, supplied by the posterior cerebral artery, do not contribute to the measurements made in the MCA. Two 2-MHz transducer probes, which are relatively insensitive to participant motion, were mounted on a screw-top headset and positioned bilaterally over the temporal skull window (Deppe et al., 2004).

Handedness
Handedness was assessed via the Edinburgh Handedness Inventory (EHI; Oldfield, 1971). The inventory consists of 10 items assessing
The Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007) was used to assess language history for all participants. The LEAP-Q is a self-assessment questionnaire consisting of nine general questions and seven additional questions per language that explore acquisition history, context of acquisition, present language use, and language preference and proficiency ratings across language domains (speaking, understanding and reading) as well as accent ratings. Rating scales either involve preference ordering (e.g. please list all the languages you know in order of dominance), stating age or specifying time in years (e.g. age when you began acquiring a language; please list the number of years and months spent in each language environment), or ratings on a scale of 0 to 10. Numeric ratings were used to assess proficiency and months spent in each language environment), or ratings on a scale of 0 to 10. Numeric ratings were used to assess proficiency and months spent in each language environment). A general measure of accent was calculated by averaging across all items (“always left” -100; “usually left” -50; “both equally” 0).

Language history questionnaire
The Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007) was used to assess language history for all participants. The LEAP-Q is a self-assessment questionnaire consisting of nine general questions and seven additional questions per language that explore acquisition history, context of acquisition, present language use, and language preference and proficiency ratings across language domains (speaking, understanding and reading) as well as accent ratings. Rating scales either involve preference ordering (e.g. please list all the languages you know in order of dominance), stating age or specifying time in years (e.g. age when you began acquiring a language; please list the number of years and months spent in each language environment), or ratings on a scale of 0 to 10. Numeric ratings were used to assess proficiency (0 “no proficiency” to 10 “perfect proficiency”), learning influences (0 “not a contributor” to 10 “most important contributor”), exposure (0 “never” to 10 “always”) and accent (0 “none” to 10 “pervasive”).

The complex range of responses to the LEAP-Q was reduced to six core measures. A general proficiency measure was obtained by averaging the scores for understanding, speaking and reading for L2. Age of acquisition for L2 was assessed by the parameter “age started learning language”. A general measure of accent was calculated averaging across “accent perceived by self” (0 “none” to 10 “pervasive”) and “accent identified by others” (0 “none” to 10 “pervasive”) for L2 (M = 3.08; SD = 2.66). The time spent in an L2 country was used as a measure of immersion for L2. For both learning of and exposure to L2, participants were asked to rate how much factors contributed to learning or are part of their current language exposure (0 “no relevance” to 10 “extreme relevance”); we used the report based on “friends”, since participants reported friends as playing an important role in L2 acquisition and exposure.

Word generation task
Tasks were programmed using Presentation® software (version 17.2; www.neurobs.com). All instructions were presented centrally in white Arial font on a black background. Each participant was tested in English (L2) and their native language (L1; French or German) in a single session using two tasks, each consisting of 23 trials.

The order of the two language tasks was counterbalanced across participants and the entire testing session lasted between 75 and 90 minutes. The experimenter spoke English at all times. So that they were focussed on their native language, participants were asked to describe the Cookie Theft picture of the Boston Diagnostic Aphasia Examination in their native language prior to being tested in that language (Goodglass & Kaplan, 1983).

The cued word generation paradigms were based on Knecht and colleagues’ 1998 paradigm (Knecht et al., 1998b). For each trial, the participant is shown a letter and is asked to silently generate words starting with that letter. Each task comprised 23 trials and lasted for around 20 minutes. We excluded the three letters with the lowest first letter word frequency: Q, X and Y in English; Q, X and Z in German; and W, X and Y in French. Task instructions for the German and French word generation tasks were translated into German and French by the experimenter.

Each trial started with an auditory tone and the written instruction “Clear Mind” (5s), followed by the letter cue to which the participant silently generated words (15s), and then overt word generation (5s) (Figure 1). To restore baseline activity, participants were instructed to relax (25s) at the end of each trial. Event markers were sent to the Multi-Dop system when the letter cue appeared, denoting trial onset for subsequent analysis of the Doppler signal.

Data pre-analysis and calculation of asymmetry indices
The blood flow data recorded from the Doppler probes were down-sampled (25 Hz) and analysed using a custom version of the analysis program developed for word generation tasks (DopOSC; Badcock et al., 2012a) running in Matlab R2014a (https://uk.mathworks.com/products/matlab/whatsnew.html; Mathworks, Natick, MA, USA). FTCD data were imported in text format with each column representing one data channel. The data were segmented into epochs, starting at baseline and ending after 30s. Trial baseline was set 10s prior to the cue tone i.e. 15s prior to letter onset (Figure 1). Following procedures developed by Deppe et al. (2004), the data were corrected for insonation angle, physiological changes unrelated to the task, and movement artefacts. The mean

![Figure 1. Word generation task.](https://uk.mathworks.com/products/matlab/whatsnew.html)
of each epoch was used to normalise the data to an arbitrary value of 100 to allow comparison across epochs, and the data were baseline corrected to eliminate slow spontaneous changes in CBFV (Deppe et al., 2004). The analysis program further included a function for heart cycle integration (Deppe et al., 2004; Badcock et al., 2012b), artefact rejection and an extra correction to interpolate the signal during very brief periods of dropout or spiking; this could be used once per trial.

Epochs were rejected when the normalised Doppler velocity was extremely high (>140 cm/s) or extremely low (< 60 cm/s), as these were considered to be measurement artefacts. Epochs were also rejected when values were more than three standard deviations below or four standard deviations above channel mean. Additionally, epochs were manually excluded when issues during the testing session had been recorded (e.g. participant speaking during rest period). The period of interest (POI) was defined between 8 and 20s after cue tone i.e. 3–15s after letter appeared (Figure 1). For each language, normalised left minus right difference values were determined. This is usually referred to as a ‘laterality index’, but here we use the term ‘asymmetry index’ to avoid confusion between the abbreviations LI (for laterality index) and L1 (for first language). Asymmetry indices were calculated as the mean left-right difference across a 2-s window centred on the maximum peak difference within the POI. Positive values indicated left lateralisation, while negative values indicated right lateralisation.

### Statistical analysis

All statistical data analyses were performed using SPSS version 22.0 (IBM Corp., Armonk, NY).

### Results

Parametric inferential statistics were used for the statistical analyses of the asymmetry indices, since they were distributed normally. Individual differences in data were analysed using non-parametric correlations. Overall, the statistical analyses consisted of analysis of variance, paired samples and independent samples t-tests, and parametric as well as non-parametric correlational analyses.

### Asymmetry indices

Normalized Doppler velocities for the left and right hemisphere activation, as well as left minus right differences, are presented for each task in Figure 2 and Figure 3. Descriptive statistics for the general Doppler data are presented in Table 1.

At the group level, both language versions of the word generation task produced significantly left-lateralised blood flow. Participants were categorised as left-lateralised if the mean asymmetry index was positive and right-lateralised if the mean asymmetry index was negative or the 95% confidence interval did not span zero. If the 95% confidence intervals crossed zero, language was considered to be bilaterally represented. There were 22 participants...
categorised as left-lateralised for both L1 and L2, three participants categorised as bilateral for L1 and left-lateralised for L2, and one participant categorised as left-lateralised for L1 and bilateral for L2 (Figure 4).

Asymmetry indices were analysed using a 2x2 mixed ANOVA with language (L1 vs. L2) as within-subjects factor and first language (German vs. French) as between-subjects factor. The means and standard errors for all dependent measures are shown in Table 2. The main effects of language and first language were not significant; for L1/L2: F (1, 24) = 1.12; p = 0.352; French/German: F (1, 24) = 1.80; p = 0.193. The interaction between these factors was also nonsignificant: F (1, 24) = 0.9; p = 0.352. Thus, there was no difference in the asymmetry indices between L1 and L2 for either French or German speakers.

Asymmetry indices for L1 and L2 were significantly positively correlated (Pearson’s: r = 0.81; p < 0.001; n = 26; Figure 4).

To check the reliability of the asymmetry index, Pearson’s product moment correlations between odd and even trials for L1 and L2 were calculated. Odd and even trials for L1 were significantly positively correlated (r = 0.74; p < 0.001; n = 26). However, odd and even trials for L2 were not significantly correlated (r = 0.28; p = 0.162; n = 26).

Independent sample t-tests were run to assess the role of testing order and compare asymmetry indices between L1 tested first and L2 tested first for L1 and L2; in neither case was the difference significant (Table 3).

Handedness
The handedness measures from the EHI ranged from -22.73 to 100 (M = 73.43; SE = 5.05). Only one participant was left-handed for writing. The handedness score was not correlated with the asymmetry index for either language, nor for the difference in asymmetry indices between languages (absolute values for Spearman correlations were all <0.15).

Language history
Self-reported measures of language history, as assessed by the LEAP-Q, are displayed in Table 4.

The data were tested for normality using the Shapiro-Wilk test. Non-normality was significant for a variety of questionnaire measures, but was not significant for either asymmetry index.

Table 5 shows Spearman’s correlations to explore the relationship between individual differences and asymmetry indices.

**Table 1. Number of included trials and asymmetry indices for L1 and L2.** Descriptive statistics for the general Doppler data are presented: the number of included trials (out of 23) and asymmetry indices for first and second language. L1, first language; L2, second language; N Trials, number of included trials; Min, minimum; Max, maximum; SEM, standard error of the mean.

| Measure     | L1          | L2          |
|-------------|-------------|-------------|
| N Trials    | Mean 22.42  | 22.08       |
|             | Min 19      | 20          |
|             | Max 23      | 23          |
| Asymmetry index | Mean 3.44   | 3.66        |
|             | SEM .39     | .28         |

![Figure 4. Scatterplot asymmetry indices L1 and L2.](image) Scatterplot displaying the relationship between asymmetry indices for L1 and L2 word generation tasks. Error bars indicate 95% confidence intervals. L1, first language; L2, second language.
The role of testing order was assessed with two independent sample t-tests, comparing asymmetry indices between L1 tested first and L2 tested first, for L1 (German or French) and L2 (English); in neither case was the difference significant. L1, first language (German or French); L2, second language (English); N, number of participants; SEM, standard error of the mean.

Table 2. Asymmetry indices (mean and SEM) for L1 and L2 word generation task separated by first language. Descriptive statistics of asymmetry indices for first and second language word generation tasks separated by first language are presented. Asymmetry indices were analysed using a 2x2 mixed ANOVA with language (L1 vs. L2) as within-subjects factor and first language (German vs. French) as between-subjects factor. Main effects and interaction effects were not significant. L1, first language; L2, second language; N, number of participants; M, mean, SEM, standard error of the mean.

| First language | N | L1 M | SEM | L2 M | SEM |
|----------------|---|------|-----|------|-----|
| German         | 15| 3.90 | .51 | 3.93 | .37 |
| French         | 11| 2.82 | .59 | 3.30 | .43 |

Table 3. Asymmetry indices (mean and SEM) for L1 and L2 word generation task, separated by order of testing. The role of testing order was assessed with two independent sample t-tests, comparing asymmetry indices between L1 tested first and L2 tested first, for L1 (German or French) and L2 (English); in neither case was the difference significant. L1, first language (German or French); L2, second language (English); N, number of participants; SEM, standard error of the mean.

| Testing Order  | N | L1 M | SEM | L2 M | SEM |
|----------------|---|------|-----|------|-----|
| L2 tested first| 12| 2.94 | .49 | 3.52 | .38 |
| L1 tested first| 14| 3.87 | .59 | 3.78 | .41 |

Table 4. Means, SEM and ranges for language history measures from LEAP-Q. Descriptive statistics of the six core measures of the LEAP-Q we selected and computed are presented. High mean score for “general proficiency” and early mean age for “age started learning second language” are evident. SEM, standard error of the mean; L2, second language; yr; years; LEAP-Q, Language Experience and Proficiency Questionnaire.

| Measure                        | Mean | SEM | Range     |
|--------------------------------|------|-----|-----------|
| General proficiency\(a\)      | 9.09 | .19 | 6.3 to 10 |
| Age started learning L2 (yr)  | 7.62 | .83 | 0 to 15   |
| Accent\(c\)                   | 3.08 | .52 | 0 to 8    |
| Time spent in L2 country (yr) | 4.93 | 1.04 | 0.3 to 21 |
| Learning (friends)\(d\)       | 8.04 | .57 | 1 to 10   |
| Exposure (friends)\(d\)       | 8.00 | .44 | 0 to 10   |

\(a\)Range: 0 (none) to 10 (perfect)
\(c\)Range: 0 (none) to 10 (pervasive)
\(d\)Range 0 (no relevance) to 10 (extreme relevance)

None of the correlations between asymmetry indices and individual differences approached significance at a Bonferroni-corrected significance level of 0.05/18 = 0.002.

Discussion
We had anticipated that language lateralisation might be stronger for a bilingual person’s first language than for the second language, especially for those who were less proficient or who learned the second language later, but this was not confirmed. Nearly all participants showed significant left lateralised blood-flow for both L1 and L2 during the word generation task; only four participants were classified as bilateral for one language, and for three of these it was L1 that was bilateral.

Furthermore, asymmetry indices for L1 and L2 were highly related and similar in magnitude, suggesting good reliability of the measure. When the two languages were considered separately, the correlation between odd and even trials was substantial for L1, but less for L2 asymmetry indices. At first glance this suggests that lateralisation may be more labile in a second language, but we suspect that a more prosaic explanation is likely, i.e. that this result reflects poor quality of the signal in a subset of participants/trials. The odd/even statistic is derived from around 10 trials per condition, and is vulnerable to the effects of outliers. If L2 lateralisation genuinely was more labile, we would not have expected to see the good agreement between the L1 and L2 asymmetry indices when all 23 trials were considered.

We also failed to find any support for the hypothesis that language proficiency affected either absolute levels of language lateralisation or the difference in lateralisation between the two languages. However, this could be due to the level of proficiency in our participants, which was generally very high. Additionally, handedness and asymmetry indices were unrelated. Although in large populations, these variables are related, the association is not strong and it is not uncommon to fail to observe it in small samples (Bishop et al., 2009).

Table 5. Spearman’s correlations between asymmetry and language history measures (\(N = 26\)). Spearman’s correlations to explore the relationship between individual differences and asymmetry indices were carried out. None of the correlations approached significance at Bonferroni-corrected significance level of 0.05/18= 0.002. L1, first language; L2, second language.

| Measure                        | Asymmetry: L1 | Asymmetry: L2 | Asymmetry difference between L1 and L2 |
|--------------------------------|---------------|---------------|---------------------------------------|
| Proficiency (general)          | -.00          | -.01          | .10                                   |
| Age (started learning)         | -.08          | -.02          | .09                                   |
| Accent (general)               | -.02          | -.01          | -.11                                  |
| Immersion (country)            | .07           | .14           | -.05                                  |
| Learning (friends)             | -.24          | .00           | .34                                   |
| Exposure (friends)             | -.35          | -.13          | .35                                   |
Limitations

1. Sample population and task type. Studies using behavioural measures of laterality have found that age of acquisition and proficiency are influential factors; early bilinguals tend to show bilateral involvement for both languages, while late bilinguals are left lateralised (Hull & Vaid, 2007). Our study appears to contradict this finding, but it should be noted that our sample only displays a limited range of age of acquisition and proficiency, raising the possibility that we may have replicated associations between asymmetry and language proficiency or age of acquisition with a more heterogeneous sample.

In addition, we used only one language task, the word generation task. This has been regarded as the gold standard task for FTCD studies, but it may not be optimal for uncovering inconsistent laterality between languages. Previous studies have found that the differential impact of age of acquisition, proficiency and exposure effects is task dependent (Dehaene & Cohen, 1997; Kim et al., 1997; Klein et al., 1995; Klein et al., 2003; Wartenburger et al., 2003). The productive word generation task may be regarded as taxing lexical-semantic processing and fMRI studies have shown that during a lexical-semantic sentence judgement task, only low proficiency individuals display greater activation in language associated areas independent of age of acquisition (Wartenburger et al., 2003). However, the literature is not consistent on this point, and other studies identify contrasting conclusions, revealing exposure dependent differences in activation during fluency tasks in high proficiency early bilinguals (Abutalebi et al., 2005; Perani et al., 2003).

In future work, laterality and multilingualism need to be examined through different tasks because both are multidimensional constructs (Bishop et al., 2013). When other tasks assessing different language components have been used with FTCD, variations in strength or direction of lateralisation became apparent in monolinguals (Badcock et al., 2012b; Bishop et al., 2009; Haag et al., 2010; Stroobant et al., 2009). The multidimensionality of laterisation, in particular for language, is suggested by the fact that individual tasks have higher split-half reliabilities than inter-correlations between asymmetry indices from different tasks (Bishop et al., 2009; Bishop, 2013).

We found no difference in laterality patterns between French-English and German-English bilinguals, but it is possible that differences might be more apparent with languages that are more different from one another, in grammatical structure, lexical items and/or phonology. These factors have been shown to influence the ease with which a second language is learned, and might plausibility affect the extent to which language representations might be shared or distinct (Schepens et al., 2016).

2. Language assessment. We used a self-report questionnaire to describe our sample and assess language history and proficiency, but behavioural measurements of proficiency may have revealed a wider response range for correlational analysis. Although Marian and colleagues established high reliability and validity for the self-report questionnaire used here, and validated it against behavioural measures, their questionnaire was devised to describe a population rather than provide an analysis measure of individual differences (Marian et al., 2007).

3. Method. While test-retest reliability of FTCD measurements is high and the time-locked correlation analysis of CBFV is robust and non-invasive, the main limitation of the method is that findings can only be interpreted on a hemispheric level, and do not give the level of localisation possible with other imaging methods (Somers et al., 2011). To clarify whether the same brain regions within a hemisphere are involved for processing first and second languages, we need measures that provide finer-grained information about within-hemisphere localisation, such as fMRI (Abutalebi & Green, 2007).

Conclusion and outlook

Participants were significantly left lateralised for both L1 and L2 when laterality was measured using FTCD during modified versions of the well validated word generation task. The high proficiency group of German-English and French-English bilinguals showed no significant difference between asymmetry indices for L1 and L2, and asymmetry indices were unrelated to language history measures and handedness. However, since laterality and language are multidimensional constructs, FTCD should be used to test bilingual laterality with different tasks and a larger, more heterogeneous sample. As an inexpensive, non-invasive, comfortable, easily applicable, mobile and child-friendly method, with a high temporal resolution, FTCD facilitates testing of large samples, tracking changes throughout development and repeated administration with different tasks. The advantages offered by FTCD are particularly interesting for research on bilinguals, which can therefore potentially outweigh its drawbacks, depending on the research question and study constraints. Furthermore, using FTCD for large-scale assessment of bilinguals could be used to interpret relationships between other cognitive functions and language. However, since FTCD cannot be used to localise specific areas related to changes in functional responses, any conclusions related to localisation have to be drawn carefully and need to be interpreted with reference to other methods.

Data availability

Open Science Framework: Bilingual FTCD, doi: 10.17605/OSF.IO/VD9DT (Bishop et al., 2016)

Please see the Data Dictionary for a description of the files.

Author contributions

Conception and design (CRG, KEW, DVMB); data collection (CRG); data processing and statistical analyses (CRG, DVMB); interpretation of data (CRG, KEW, DVMB); drafting and revising the manuscript (CRG, KEW, DVMB).

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by the Wellcome Trust programme [082498].
Open Peer Review

Current Peer Review Status: 

Version 1

Reviewer Report 20 March 2017

https://doi.org/10.21956/wellcomeopenres.10640.r20237

© 2017 Jansen A et al. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Andreas Jansen
Laboratory for Multimodal Neuroimaging (LMN), Department of Psychiatry, University of Marburg, Marburg, Germany

Verena Schuster
Department of Psychiatry, University of Marburg, Marburg, Germany

Although functional asymmetries between the hemispheres have been known since the mid-19th century, we still lack a thorough understanding of the underlying mechanisms. In particular, we do not have precise models that reveal which factors drive hemispheric specialization, how lateralization processes of different cognitive functions interact with each other, and how the brain integrates processes that are lateralized to opposite hemispheres. In the present study, Grabitz and colleagues aimed to investigate whether the hemispheric lateralization of first and second languages is different. Hemispheric dominance was assessed by functional transcranial Doppler sonography (fTCD) in 26 high proficiency bilinguals with either German or French as their first language (L1) and English as their second language (L2). fTCD was used to assess task-dependent blood flow velocity changes in the left and right middle cerebral arteries during a cued word generation task. The authors report that the majority of participants (22/26) were significantly left lateralized for both L1 and L2. They found no significant difference between the lateralization of L1 and L2, as assessed by a lateralization index (LI). They conclude that in highly proficient bilinguals, there is strong concordance for cerebral lateralization of first and second languages. Although the study was competently performed, there are some concerns about the conceptual planning of the study and the application of fTCD.

Conceptual foundations of the study: There are many aspects of functional neuroanatomy that might differ between L1 and L2, for instance the recruitment of brain regions, the strength of brain activity in specific regions or the connectivity between language regions. Hemispheric lateralization is only one aspect. It might have been useful to explain why the authors assessed in particular hemispheric dominance, it might have been useful to state why they anticipated that language lateralization is stronger for a bilingual person's first language than for the second language, and it might be have been useful to explain what they authors would have had concluded when they had found significant differences between the lateralization of L1 and L2 –
expect that there are significant differences. To interpret non-significant differences, as in the present study, it is also necessary to explicitly state how strong the LI would be expected to differ between L1 and L2. What is a minimal difference that would have been considered as relevant? It is also not clear whether the authors intended to assess differences between L1 and L2 on a group level or in individual subjects? What would be the putative role of interindividual differences? In summary, in its present version of the manuscript a theoretical concept is completely missing. Without this concept, it is not possible to properly interpret the findings. The manuscript gives the impression that the authors were just looking for differences in a rather exploratory way.

**Application of fTCD:** Before performing a study, it might be a useful exercise to ask whether the imaging technique used is a suitable tool to answer the question asked. I have serious doubts that fTCD can be applied for that purpose. The authors expect to find differences between the lateralization of L1 and L2. It is important to know whether the technique is sufficiently sensitive to find differences, if they exist. As mentioned before, the authors do not explicitly state what differences they expect. In our opinion, it is rather unlikely that the hemispheric dominance (left, right, bilateral) of L1 and L2 will be different. If a subject is for instance left dominant for L1, we do not expect that she will be right dominant for L2. The expected differences will most likely be on a smaller scale. A subject that is left dominant for L1 might be a bit less left dominant for L2. Is fTCD able to find these differences? Unfortunately, there are no methodological studies that assessed how sensitive fTCD is to find potentially small differences in the degree of lateralization. We certainly agree that fTCD is a useful tool to determine hemispheric dominance (that is, left- or right-hemispheric lateralization). It is, however, unknown if the technique can be used to also assess small differences in the degree of hemispheric lateralization. Large methodological studies in this regard, in particular from independent groups (i.e., not from the developers of AVERAGE), are missing. One might also ask why the developers report correlations between fTCD and other techniques (such as fMRI or the Wada test) as high as r~0.9 (and even much higher), when it is not possible to reproduce these findings even with the same modality. Furthermore, fTCD assesses blood flow velocity changes in the vascular territory of the left and right middle cerebral artery. This territory however shows a high interindividual variability. While one might argue that main network nodes of the language system, such as “Broca’s area”, lie within this territory in all subjects, other regions that are also active during the task might be included in the calculation of the LI in some subjects, but not in others. What are the consequences when one compares a LI between subjects? In summary, it is unclear whether fTCD is sensitive enough to measure small differences between the lateralization of L1 and L2.

To conclude, the study deals with an interesting topic and is competently performed. However, the theoretical foundation should be described in more detail, the expected difference between the LI of L1 and L2 should be reported, and it should be made clear that fTCD is able to measure the expected differences at all.

**Competing Interests:** No competing interests were disclosed.

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.**
Dorothy Bishop, University of Oxford, Oxford, UK

‘explain why the authors assessed in particular hemispheric dominance’ – many reasons for differences in L1 and L2: recruitment of brain regions, strength of brain activity, connectivity between language regions

We now make it clear that we recognise that there are potentially many ways in which language processing may differ for the two languages in bilinguals, but we do not think that invalidates a decision to look specifically at brain lateralisation, which has previously been discussed as potentially differing between languages.

State why it was anticipated that language lateralization is stronger for a bilingual persons L1 than for L2. Explain what the authors would have concluded when they had found significant differences between the lateralisation of L1 and L2. Explicitly state how strong the L1 would expected to differ between L1 and L2.

We now go into more detail regarding predictions from prior literature. The prediction of discrepant laterality between languages was not strong: In the literature, there are reports of both the same strength of lateralisation for L1 and L2 and also reduced lateralisation for L2. A finding of significant difference in lateralisation between L1 and L2 would have lent further support to one side of this debate.

What is a minimal difference that would have been considered as relevant? There are issues with interpreting non-significant findings.

As well as reporting Bayes Factors for mean comparisons, we have now conducted further analysis using the Bland-Altman method, which is specifically designed to address this issue.

It is also not clear whether the authors intended to assess differences between L1 and L2 on a group level or in individual subjects? What would be the putative role of interindividual differences?

This is a within-subjects study, with each person tested in both their languages, so the differences are evaluated in individual subjects. The correlations that are reported depend on there being individual differences in the extent of lateralisation. The result, therefore, hinges on interindividual differences.

Important to know if technique is sensitive to find differences if they exist – are there methodological studies to assess the sensitivity of ftCD to small differences in lateralisation? 'I have serious doubts that ftCD can be applied for that purpose. The authors expect to find differences between the lateralization of L1 and L2. It is important to know whether the technique is sufficiently sensitive to find differences, if they exist.'

Since this study was conducted, we have reported a study of test-retest reliability of laterality indices assessed using ftCD, which we now cite. They are high enough to give confidence that the degree, as well as direction of laterality measured this way, is reasonably stable. See Woodhead, Z. V. J., Bradshaw, A. R., Wilson, A. C., Thompson, P. A., &
Bishop, D. V. M. (2019). Testing the unitary theory of language lateralization using functional transcranial Doppler sonography in adults. *Royal Society Open Science, 6*(3), 181801. https://doi.org/10.1098/rsos.181801.

The reviewers clearly have a very negative impression of fTCD as a measure of laterality, but it's unclear what they prefer. The Wada technique is a blunt instrument that is useful in clinical contexts for making a basic distinction between left, right and bilateral, but it is neither feasible nor useful for measuring degrees of lateralisation. With fMRI one can quantify the LI, but the results will depend on the statistical approach (e.g. height or extent of statistic, %signal change etc), ROI studied and on thresholding. The kinds of individual difference in vasculature that the reviewers mentioned may well affect the observed LI - we now make that point in the Discussion. However, this will be as true for measures from fMRI as for fTCD, and in addition, with fMRI, the issue is complicated by the possibility of individual differences in localisation of language regions.

So, while we accept that fTCD is not perfect, neither are other methods, and part of our goal in ongoing research is to use them as complementary methods. Indeed we regard it as a worthwhile endeavour in future to consider how far the LI in fTCD relates to anatomical variation. But we don't see any of these as reasons to dispense with the results we have obtained, which we regard as part of a complex pattern of evidence on these issues.

*Why did the developers report correlations between fTCD and other techniques (such as fMRI or Wada test) as high as r~0.9, when it is not possible to reproduce these findings even with the same modality?*

We cannot say why the Münster group who developed fTCD reported these correlations. Our work is independent of theirs and we have not used the Average software for some years, though the processing steps we adopt are largely the same. The correlations they originally reported were based on small sample sizes and would have large confidence intervals around them. In addition, language laterality, as conventionally measured, is usually not normally distributed and should be evaluated with a nonparametric correlation coefficient. We hope to obtain data on larger samples in future that will provide more solid evidence on the relationship between laterisation as assessed by fTCD and fMRI.

**Competing Interests:** No competing interests were disclosed.
This is a succinctly written paper reporting the novel results from a non-invasive technique (functional transcranial doppler ultrasound) that examines changes in blood flow velocities in the left and right middle arteries in response to a cued word production task in a person's native language (L1, either French or German) and in their second language (L2, English). The participants were young proficient bilingual speakers immersed in an English context. The aim was to examine the degree of lateralisation in response to this task in L1 and in L2. The data are appropriately analysed with suitable correction for the number of comparisons made where required.

**Rationale**

It is important to deploy non-invasive methods that can be used to assess brain response for a particular tasks in children and in adults. The specific question addressed concerns the extent to which L1 and L2 reveal a comparable pattern of asymmetry as revealed by the measure of blood flow velocity.

It is worth noting that both hemisphere play a role in speech processing in monolingual speakers. Functional imaging data are consistent with the idea that regional activation during speech production is bilateral for motor, premotor, subcortical, and superior temporal regions whereas middle frontal activation is predominantly left lateralised (Price, 2010). As the authors correctly note, neuroimaging data strongly implicate common regions in the processing of L1 and L2. Indeed from a neurocomputational point of view, there is no reason to envisage that the processing of a second language would recruit radically distinct regions (Green, 2003). Instead, different languages may recruit different microcircuits within common regions (e.g., Paradis, 2004). We should then expect differences attributable to the distinct phonological and syntactic properties of words in different languages and commonalities in terms of their reference to common entities. Consistent with this possibility, Correia et al. (2014), using multi-voxel pattern analysis, reported discriminating neural response in multiple temporal, parietal and frontal cortical regions to individual spoken animal nouns (horse/duck) in English and Dutch combined with an invariant response pattern to the translation equivalents (paard/eend) indicative of access to common semantic/conceptual knowledge in regions such as the anterior temporal pole. In modelling recovery post-stroke, we found that models implicating the same brain regions were equally predictive for both monolingual and bilingual speakers displaying parallel recovery patterns (Hope et al., 2015). Evidence for selective recovery post-stroke does not contradict this position, but rather points to a difficulty in control (Green, 2008). Detailed determination of this possibility in the context of speech production awaits future research. However, the Wada test (using injection of intracarotid amobarbital), referred to by the authors as the gold standard in determining lateralization, strongly implicates left hemisphere representation for both languages of a bilingual speaker (e.g., Rapport, Tan & Whitaker, 1983). A non-invasive method as reported here provides a useful adjunct despite its noted limitations in terms of identifying the microcircuits involved.

**Participant information**

Self-reported proficiency does generally correlate reasonably well with more objective measures as the authors note. Nevertheless, it is usually desirable to report such objective measures. For
instance, for vocabulary measures the various tests under the rubric of LexTale offers a good source (Brysbaert, 2013; Lemhöfer & Broersma, 2012). There are also Quick Placement tests to assess syntactic knowledge.

**Procedure**
Given the experimenter spoke English all the time how was the transition to the word generation task managed, in particular the switch from describing a picture in L1 to the naming task?

**Data analysis**
The word generation task involved an interval for the silent generation of words in response to a cued letter (15 secs) followed by a 5 second recall interval. Although this interval is short and so constrains information on relative difficulty, it is of interest to know the mean scores and their variance. If there is variance, does such variance have detectable effects on the signal?

**Estimates of reliability**
The authors nicely use odd-even trials to estimate signal reliability for the asymmetry index. This estimate proved significant for the production task in the native language (L1) but not for the second language, English (L2). If there is no asymmetry difference then shouldn't there be a significant correlation when alternate trials are taken from different language runs?

**Competing Interests:** No competing interests were disclosed.

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

---

**Author Response 24 Jun 2021**

**Dorothy Bishop,** University of Oxford, Oxford, UK

**Comments under Rationale**
We thank the reviewers for providing an overview of the literature on microcircuitry, which we now mention in the Discussion

**Usually desirable to report objective measures of language proficiency**
Please see response to reviewer 1. We did not have such measures for Study 1, but we do for study 2.

*Given the experimenter spoke English all the time, how was the transition to the word generation task managed, in particular the switch from describing a picture in L1 to the naming task?*  
Alas, this was not noted at the time of data collection for Study 1 and we do not have a record of how this was handled, though we do state that the examiner used English throughout. For Study 2, the experimenter switched instruction languages according to the tested languages.

**Data analysis.** The word generation task involved an interval for the silent generation of words in response to a cued letter (15 secs) followed by a 5 second recall interval. Although this interval is short and so constrains information on relative difficulty, it is of interest to know the mean
scores and their variance. If there is variance, does such variance have detectable effects on the signal?

We were not sure we had interpreted this correctly; in Study 1 words were generated covertly, therefore we did not have a record of responses. However, in previous studies with monolinguals, we have specifically considered whether varying task difficulty affects laterality. Where difficulty is varied by constraining the task (requiring words starting with 2 specific letters rather than one), this reduced performance but did not affect the LI. (Badcock, N. A., Nye, A., & Bishop, D. V. M. (2012). Using functional transcranial Doppler ultrasonography to assess language lateralisation: Influence of task and difficulty level. Laterality, 17(6), 694–710. https://doi.org/10.1080/1357650X.2011.615128). In Study 2 subjects generated words overtly and we report data on number of words produced. There was no relationship between number of words generated and LI.

**Estimates of reliability** The authors nicely use odd-even trials to estimate signal reliability for the asymmetry index. This estimate proved significant for the production task in the native language (L1) but not for the second language, English (L2). If there is no asymmetry difference then shouldn’t there be a significant correlation when alternate trials are taken from different language runs?

We were also puzzled by the differing estimates of split half reliability - as it turns out when we reanalysed the data for this version, using our current analysis scripts, the estimate of split half reliability was more similar for the two languages: for L2, the original analysis gave r = .28. With our new method, one participant met criteria as an outlier and was excluded, and we also used Spearman rather than Pearson correlation, and based the LI on the mean rather than peak of the difference waveform; this gives r = .60. Please note: the analytic decisions leading to these changes were made a priori: we used the scripts and outlier exclusion criteria that we documented in Woodhead et al (2019), and list here how each modification of the method affected the correlation:
- Discarding one participant with noisy data (participant 14), R = 0.44
- Using Spearman's correlations instead of Pearson's, R=0.49
- Using mean LI method instead of peak, R=0.60

We feel this provides further justification for basing analyses on mean rather than peak values: the latter can be more noisy, especially if the data do not show a single pronounced peak.

**Competing Interests:** No competing interests were disclosed.
It pains me to have to write this review. The research done is good and reliable (and hence the Wellcome Trust may decide to publish it), but the question addressed is futile and the methods used far from optimal. Therefore, I fear that if this article is indexed to PubMed Central, it will not do the authors much good.

For a start, the authors had anticipated that language lateralization might be stronger for a bilingual’s first language than for the second language. In 1992, Paradis already called this the Loch Ness Monster of research on laterality and bilingualism. There is no sound evidence whatsoever that L2 processing would be less lateralized than L1 processing (as the authors indeed found). There is even very good evidence that as L2 proficiency increases, it increasingly uses the very same brain areas as L1 processing. Only at low levels of L2 proficiency can one sometimes see extra right and left hemisphere activity, arguably because the participants are using all types of strategies, including non-language ones.

Second, the authors are using the crudest neuroscientific technique available, fTCD. As they say, it is cheap, it can be applied easily (but leads to a considerable loss of participants), but it is also very crude, as it only compares to blood flow to the left vs. the right hemisphere. In the present study, the reliability is good (except for L2 processing), but even so it remains a technique that only can tell you something about more left than right processing, nothing more. So, in the end the authors are investigating a strawman hypothesis with an unformative technique.

Third, there are power issues. A lot of subject-related variables are tested on a group of 26 participants. Luckily the authors did not find anything significant, because any significance they would have found, would have been very likely due to a statistical fluke, which cannot be replicated (see papers by Gelman).

Fourth, individual differences are thought to be of interest. Still, they are studied with subjective scales. Why not measure the proficiency with a vocabulary test (e.g., LexTALE, Shipley)? Why use Likert scales? In several studies (involving the French and Spanish LexTale tests), I’ve reported that although there is a good correlation between subjective estimates on the basis of a Likert scale and the LexTale scores, for individual participants there can be a big difference, because participants use different comparison groups (e.g., L2 learners compare their performance to other L2 learners, not to L1 speakers). If the authors want to keep on using subjective measures, they may want to try descriptions as the levels defined by the European framework.

As said, if the goal of the Wellcome Open Research initiative is to make all reliable empirical data available, I am not against publication. However, for the above reasons I do not think this publication will do the authors (nor the journal) much good. The only bit of information I found of value was Figure 4. Even then it would be good to see this supported by fMRI validation. I know the Knecht group did so, but still I’d like to see it done in other groups as well. I’m very curious, for instance, to what extent the bilateral patterns are valid. We rarely see them in fMRI research.

**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, however I have significant reservations, as outlined above.

Author Response 24 Jun 2021

Dorothy Bishop, University of Oxford, Oxford, UK

Brysbaert, while unenthusiastic about our original paper, accepted that "if the goal of the Wellcome Open Research initiative is to make all reliable empirical data available, I am not against publication." Nevertheless, he queried whether the study was worth doing, given that prior work with fMRI had not found discordant laterality for two languages in bilinguals. In addition, it was felt that reliance on self-reported proficiency was non-optimal, and that there were concerns about concluding a lack of difference between languages when sample size, and hence statistical power, was low. He also expressed misgivings about the method we had used, functional transcranial Doppler ultrasound.

1. **Prediction of different laterality in L1 and L2 is futile. We already know that is not the case.**

If we understood this point correctly, the reviewer is arguing that the fact that laterality is the same in L1 and L2 for proficient bilinguals is so well established that it is pointless to provide a further demonstration of the point. We disagree. The literature has not always been consistent and most studies are small, so an accurate picture may only become clear when there is sufficient information for a meta-analysis. Our aim was to use fTCD to contribute to this literature. We accept that the reviewer has a very low opinion of fTCD, but we do not think this is justified, and indeed would argue that the strong correlations between L1 and L2 laterality indices obtained with this method provide some evidence that individual variation in degree of lateralisation are meaningful.

2. **fTCD is not sensitive enough to detect relevant hemispheric differences**

We now provide more arguments in support of fTCD. We note also reviewer 2's comment: 'A non-invasive method as reported here provides a useful adjunct despite its noted limitations in terms of identifying the microcircuits involved.' Brysbaert states that fTCD only tells you about left- vs right hemisphere blood flow. That's exactly what we are interested in, so this criticism does not seem valid. We should stress we are not making massive claims for fTCD - it clearly has its limitations - but dismissing a study on laterality just because it uses this method seems premature. It is one tool in the range of possible methods: we need to do more work with all of them (behavioural, fTCD, fMRI) to study how they relate to one another and how reliable and sensitive they are, in order to make progress in laterality research. This is exactly what we are doing in our current research programme.

3. **there are power issues. A lot of subject-related variables are tested on a group of 26 participants.**

We agree that there are power issues. Brysbaert claims the result is unlikely to replicate. We have now included Study 2 - this confirms that the result does replicate, and generalises to another language and task (semantic fluency).
In terms of the subsequent exploratory analysis of correlations with other measures, as noted by reviewer 2: 'The data are appropriately analysed with suitable correction for the number of comparisons made where required.' (our emphasis). However, we agree the sample is too small for sensible exploratory analyses, and have now modified our focus to the Age of Acquisition effect, which is a matter of some debate in the literature. Other variables are reported for completeness, but we agree that it is not sensible to report all correlations in the absence of a priori predictions.

4. ‘individual differences are thought to be of interest. Still, they are studied with subjective scales. Why not measure the proficiency with a vocabulary test (e.g., LexTALE, Shipley)? Why use Likert scales? In several studies (involving the French and Spanish LexTale tests), I’ve reported that although there is a good correlation between subjective estimates on the basis of a Likert scale and the LexTale scores, for individual participants there can be a big difference, because participants use different comparison groups (e.g., L2 learners compare their performance to other L2 learners, not to L1 speakers). If the authors want to keep on using subjective measures, they may want to try descriptions as the levels defined by the European framework’

We agree. This point was also made by Reviewers 2, though they also pointed out that self-reported proficiency correlates reasonably well with more objective measures. A strength of study 2 is that it used more objective measures of language proficiency. We thank the reviewer for the excellent suggestions for measures to be used in future work.

**Competing Interests:** No competing interests were disclosed.