Bilingualism Results

Sophie Harte/Zoe Woodhead/Dorothy Bishop

9th April 2021

## Warning: NAs introduced by coercion

# Study 1: French/German Speakers

## Methods

### 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). Data was collected from 26 participants. During the analysis,1 further participant was dropped because of an insufficient number of useable trials. All further analyses are based on the final sample of 25 participants (19 female; mean age = 22.84 years, sd = 3.7 years).

### 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.

### Handedness

Handedness was assessed via the Edinburgh Handedness Inventory (EHI; Oldfield, 1971). The inventory consists of 10 items assessing dominance of a person’s right or left hand in everyday activities. Each item is scored on a 5 step scale (“always left”, “usually left”, “both equally”, “usually right”, “always right”). A person can score between -100 and +100 for each item and an overall score is calculated by averaging across all items (“always left” -100; “usually left” -50; “both equally” 0).

### Language History

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. An overall self-rerported proficiency rating was calculated by taking the mean ratings for proficiency in speaking, reading and understanding English.

The main variable of interest from LEAP was age of acquisition of L2 (AoA); following Hull and Vaid (2007), we subdivided into early AoA (before 6 years of age) and late AoA subgroups, to test the prediction that language is more bilaterally represented when L2 is learned in infancy. To characterise the sample, we also report the numbers of languages spoken; age of achieving fluency in English; self-reported strength of foreign accent when speaking English (on a scale from 0 [none] to 10 [pervasive]); and mean self-reported proficiency in English.

### fTCD 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 cerebral blood flow velocity (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).

### 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 languages 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.

Note that, in contrast to Experiment 2, the participants’ responses on the Word Generation task were not recorded.

### fTCD Analysis

The CBFV data were analysed using custom scripts in R Studio (R Studio Team, 2015), which are available on Open Science Framework (link). The data preprocessing followed conventional methods (Deppe, Ringelstein & Knecht, 2004), and included the following steps:

* Downsampling from 100 Hz to 25 Hz
* Epoching from -11 s to 21 s relative to the onset of the ‘Clear Mind’ cue
* Manual exclusion of trials with obvious spiking or dropout artefacts
* Automated detection of data points with signal intensity beyond 0.0001-0.999 quantiles. If a trial contained one of these extreme data points, it was replaced by the mean for that epoch; if it contained more than one, the trial was excluded from further analysis
* Normalisation of signal intensity by dividing CBFV values by the mean for all included trials and multiplying by 100
* Heart cycle integration by averaging the signal intensity from peak to peak of the heart beat
* Baseline correction by subtracting the mean CBFV across the baseline period (-10s to 0s relative to the ‘Clear Mind’ cue) from all values in the trial
* Automated detection and rejection of trials containing normalized values below 60 or 140.

Participants with less than 15 usable trials for either language were excluded from all further analyses. For each participant that was included in the analysis, a grand mean was calculated over all of their included trials. A laterality index (LI) was calculated by taking the mean of the difference between left and right CBFVs (L-R) within a period of interest (POI) that started 8s after the ‘Clear Mind’ cue (i.e. 3s after the word generation task had begun) and ended at 20s (i.e. when the covert generation task ended). The start time of the POI was chosen to allow time for the blood flow to respond to the task; and the end time was chosen to prevent capturing the response to the overt speech generation phase.

This method of calculating LI using the mean L-R difference across the whole of the POI (the ‘mean’ method) deviates from the conventional method that we had used in the first version of this paper. The original ‘peak’ method, popularised by Deppe, Knecht, Henningsen and Ringelstein (1997) takes the mean of a narrow time window around the peak difference within the POI. This method forces the LI to be either left or right - even if the waveform is close to zero with no clear lateralised peak, the highest absolute value in the POI will be treated as a peak. This creates a bimodal distribution of LIs. We have compared the ‘peak’ method with our ‘mean’ method, and shown that, while they give high agreement, the mean method gives reliable measures and normally distributed LI values, albeit with lower values, due to averaging over the whole POI (Woodhead et al., 2020). We have therefore moved to using the mean method in our current research. Nonetheless, peak LI values were computed in case they are required for comparison with other studies, and are available on the online data repository: <https://osf.io/4pm76/>.

Finally, following Woodhead et al (2019), we identified and excluded datasets with unusually high trial-by-trial variability using the Hoaglin and Iglewicz (1987) outlier detection method. For this analysis, LI was calculated for each trial, rather than just for the grand average. The standard error of the LI values was then calculated. Outliers were defined as datasets where the standard error was above an upper threshold, calculated as:

Upper threshold = Q3 + 2.2 \* (Q3 – Q1)

whereby Q1 is the first quantile of the standard errors among all participants, and Q3 is the third quartile. Participants who had standard error above the upper threshold for either L1 or L2 were excluded from all further analyses.

### Statistical Analysis

To test our main hypothesis, the association between strength of lateralization (LI values) for L1 and L2 was first visualized using a scatterplot, with the strength of association computed as Spearman’s correlation coefficient. Following Woodhead, Rutherford, and Bishop (2020), we adopted an approach based on Bland and Altman (1986) to determine whether the LIs for L1 and L2 were equivalent. This involves specifying boundaries for the expected distribution of difference scores, assuming that the two values are equivalent. Woodhead et al (2020) computed difference scores by LIs for odd vs even trials, and set boundaries corresponding to expected mean of zero +/-1.96 standard deviations. If the two measures were equivalent, 95% of difference scores between LIs for L1 and L2 should fall in this range (from -2.5 to 2.5).

For our second hypothesis, we compared laterality for L2 between those with early vs late Age of Acquisition, using a one-tailed test of the prediction that early AoA would be associated with lower LI for L2.

In addition, we report the correlation between LI values and strength of handedness (EHI quotient), and the impact of testing order (L1 then L2, or L2 then L1).

## Warning in cor.test.default(background\_data$EHI, L1\_data$L1.mean\_LI, method =  
## "spearman"): Cannot compute exact p-value with ties

## Warning in cor.test.default(background\_data$EHI, L2\_data$L2.mean\_LI, method =  
## "spearman"): Cannot compute exact p-value with ties

## Results

### Handedness

Summary statistics for the EHI handedness measure can be seen in Table 1. All participants but one had EHI values above 0, indicating right handedness. The remaining participant had an EHI of -20, indicating weak left handedness. Correlations between LI from fTCD and handedness scores on the EHI, were not significant for either L1 (r = -0.041) or L2 (r = 0.078).

Table 1. Demographics for the Study 1 participants, N=25 (19 female)

| Characteristic | Mean (sd) |
| --- | --- |
| Age, years | 22.84 (3.7) |
| EHI/100 | 73.27 (26.25) |
| Languages spoken | 3.64 (0.99) |
| Age of English acquisition, years | 7.56 (4.32) |
| Age of English fluency, years | 12.08 (6.7) |
| English accent/10 | 2.52 (2.38) |
| English overall score/10 | 9.09 (0.98) |
| English speaking/10 | 8.92 (1.08) |
| English listening/10 | 9.12 (1.05) |
| English reading/10 | 9.24 (0.93) |

### Language History

Summary statistics for the language history questionnaire can be seen in Table 1. Self-reported proficiency in speaking, reading and understanding English were all generally high (all around 9/10), with a minimum for any individual rating of 6/10. Age of acquisition was more variable, ranging from 0 to 15 years. Binary categorisation of AoA, using Hull and Vaid’s (2006) criteria gave 7 cases of early AoA (below 6 years of age), and 18 cases of late AoA.

### fTCD Data Quality and Reliability

As mentioned in the Methods, one participant was excluded from the analysis because of insufficient number of usable trials. For the remaining participants, 5.22% of trials were excluded for L1, and 5.57% for L2.

Normality of the LI values was assessed using Shapiro-Wilk tests. Distributions of LI were unimodal for both L1 and L2, and for L1 the data did not significantly deviate from normality (W = 0.89, p = 0.009). Data for L2 were significantly non-normal (W = 0.97, p = 0.577), showing a rightward skew.

Split-half reliability was assessed by correlating the LI values from odd and even trials. The Spearman’s correlation for the L1 data was 0.74, and for the L2 data it was 0.6, indicating medium to good within-session reliability.

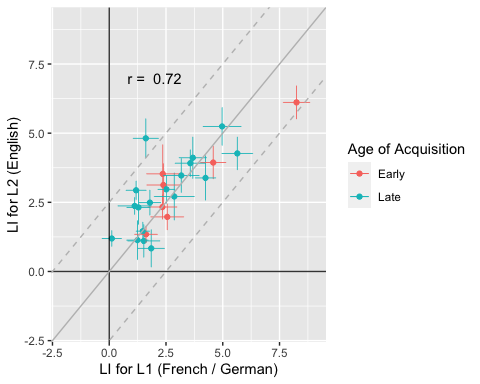
### LI Values

Table 2. Summary statistics for Study 1 laterality indices

| Language | Mean trials | mean LI | se LI | % left | % bilateral | % right |
| --- | --- | --- | --- | --- | --- | --- |
| L1 | 21.8 | 2.72 | 0.35 | 92 | 8 | 0 |
| L2 | 21.72 | 2.92 | 0.28 | 88 | 12 | 0 |

Table 2 shows summary statistics for the LI values for L1 and L2. The percentage of participants in each group categorised as left lateralised, bilateral or right lateralised is also shown. The majority of participants were left lateralised, with only around 10% showing bilateral activation. No participants showed right lateralisation for either L1 or L2. T-tests showed that there were no significant effects of testing order on LI values, either for L1 (p = 0.231) or L2 (p = 0.933).

As can be seen in the scatterplot in Figure 1, laterality indices for L1 and L2 were similar, with Spearman’s R = 0.72. Furthermore, the points cluster around the bold grey line, which shows the point of equivalence between L1 and L2, and all but one point falls within the Bland-Altman bounds (dotted grey lines), indicating that L1 and L2 are equivalent.

 **Figure 1. Scatterplot showing individual mean LIs in L1 and L2, with horizontal and vertical denoting standard errors. The solid grey line corresponds to the point of equality of the two measures, and the dotted lines show the limits where difference between LIs is +/- 2.5.**

### Effect of Age of Acquisition

One can see by inspection of Figure 1 that there is no evidence of a trend for lower LI for L2 in those with early AoA, and a t-test of differences in L2 LI for those with early and late AoA revealed no differences: t = 0.56, p = 0.59. For a more quantitative assessment of association, we computed Spearman’s correlations between the LI values for L2 (English) and the age of acquisition of English. This was not statistically significant (r = 0.03, p = 0.9).

## Discussion

All high proficiency Good agreement - validates method AoA no effect.

TO BE COMPLETED

# Study 2: Japanese Speakers

## Methods

### Participants

Participants were native speakers of Japanese, who were advanced users of English from the London area, and had lived in English-speaking countries for at least 3 years. All were self-reported right-handers, and none reported any reading or language difficulties.

A total of 25 participants were included in the final analysis, after excluding 1 participant due to poor quality data or insufficient trial numbers. All further analyses are based on the final sample of 25 participants (mean age = 29.75 years, sd = 6.52 years).

Gender data was available for 15 participants, of which, 11 were female.

### Ethics Statement

!! NEED ETHICS INFO !!

### Language History and Ability

Age of acquisition of English and number of years of using English were evaluated via self-report. As with Study 1, a binary age of acquisition (AoA) variable was created by subdividing participants into early (below 6 years) and late (6 years or over) subgroups.

English language ability was measured using the Quick Placement Test (Cambridge Local Examinations Syndicate, 2001), which assesses English reading, listening, and grammar. The test is scored out of 60. Those who scored under 40 were classed as having basic level proficiency (N = 2); between 40 and 48 were classed as having intermediate level proficiency (N = 2); and above 48 were classed as having advanced level proficiency (N = 11)

### fTCD Apparatus

The apparatus was identical to that used in Study 1.

![possibly a different headset?]

### Word Generation Task

Participants completed both phonological and semantic word generation tasks in English and Japanese, with order counterbalanced across participants Unlike in study 1, there was no silent interval for covert word generation: participants spoke the words as they thought of them. For each trial, participants saw “Clear Mind” presented on the screen for 3 seconds. The cue stimulus was then presented, and participants had 17 seconds to overtly generate as many words as possible. Participants were then instructed to relax for 16 seconds to restore baseline activity. Each trial lasted a total of 36 seconds.

### Stimuli

#### Phonological Word Generation - Japanese

Participants were presented with a cue in Kana, a Japanese phonological script. Following the Japanese mora frequency analysis conducted by Dan et al. (2012) based on the familiarity ratings in Amano & Kondo (1999), 10 of the 12 most frequent moras that are positioned at the beginning of words were selected. The two moras omitted were は (/ha/) and じ (/ji/). は was omitted because it would be pronounced /wa/ when it was the subject-marker and じ was omitted because it was the voiced sound of し (/shi/) that was included in the stimuli. Participants had to produce as many words as possible that began with the specified kana. Each kana was presented twice, and the 20 trials were presented in a pseudo-randomised order to ensure all 10 cues had been presented once before a cue was repeated.

#### Phonological Word Generation - English

Participants were presented with 10 alphabetic letters (A, B, C, F, H, M, O, S, T, W). Participants had to produce as many words as possible that began with the specified letter. The task consisted of 20 trials which were presented in a pseudo-randomised order to ensure all 10 cues had been presented once before a cue was repeated.

#### Semantic Word Generation - Japanese

10 Japanese words representing semantic categories were presented: farm animals, zoo animals, vegetables, fruits, drinks, colours, sports, pets, tools, and transport. Participants had to report as many words that matched these categories as possible. Each category was repeated twice in the semantic fluency blocks. Categories were presented in a pseudo randomised order.

#### Semantic Word Generation - English

The same semantic categories were presented in English. Participants had to report as many words that matched these categories as possible. Each category was repeated twice in the semantic fluency blocks. Categories were presented in a pseudo-randomised order.

### fTCD Analysis

## Joining, by = "ID"

The same fTCD analysis method was used as in Study 1, except that the epoch lengths were changed to match timings for Study 2. The POI started at 6s after the onset of the ‘Clear Mind’ stimulus (i.e., 3s after the word generation task had begun) and ended at 20s (i.e., at the end of the word generation task).

## Results

Table 3. Demographics for the Study 2 participants, N=25 (11 female)

| Characteristic | Mean (sd) |
| --- | --- |
| Age, years | 29.75 (6.52) |
| Age of English acquisition, years | 10.2 (3.76) |
| Time using English, years | 10.85 (4.72) |
| English overall score/60 | 68.25 (14.23) |
| English speaking/100 | 66.67 (19.24) |
| English listening/100 | 69 (16.92) |
| English reading/100 | 68.33 (16.55) |
| English writing/100 | 69 (20.81) |

### Language History

Summary statistics of language history can be seen in Table 3. Age of English acquisition ranged from 4 to 13 years.

The mean number of words produced per trial in the phonological conditions was 5.81 (sd = 1.48) for Japanese and 6.13 (sd = 1.19) for English. The mean number of words produced per trial in the semantic condition was 7.52 (sd = 1.26) for Japanese and 7.07 (sd = 1.27) for English. There was no significant difference between the mean number of words produced per trial for L1 and L2 in the phonological condition (t(26.7) = -0.65, p = 0.523) or the semantic condition (t(28) = 0.96, p = 0.344).

### fTCD Data Quality and Reliability

Normality of LI values was assessed using Shapiro-Wilk tests. For phonological tasks, data was normally distributed for L1 (W = 0.96, p = 0.504 ) and L2 (W = 0.95, p = 0.275 ). Data was also normally distributed for semantic tasks for L1 (W = 0.97, p = 0.632 ) and L2 (W = 0.98, p = 0.945).

Split-half reliability was assessed by correlating the LI values from odd and even trials, using Spearman’s correlations for consistency with Study 1. For phonological word generation, the split-half correlation was 0.6 for L1 and 0.84 for L2. For semantic tasks, the correlation was 0.61 for L1 and 0.69 for L2. This indicated moderate to good reliability for all tasks.

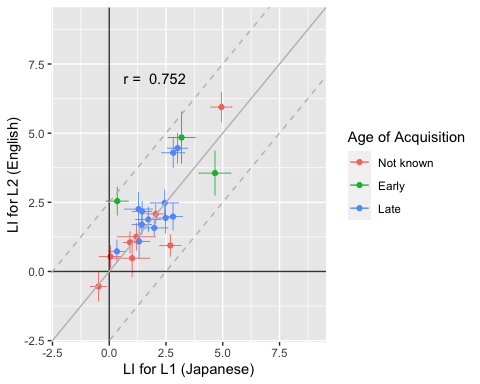
### LI Values

Table 4 shows summary statistics for L1 and L2 in both phonological and semantic word generation tasks.

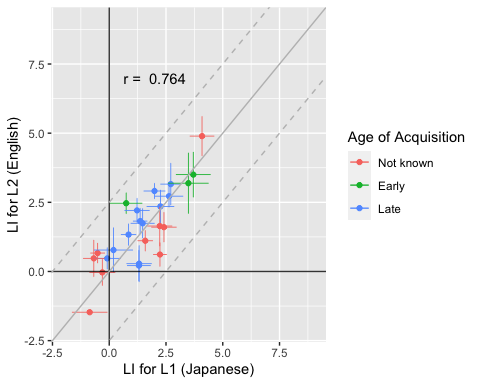
Laterality indices for L1 and L2 were strongly correlated in both the phonological task (Spearman’s R = 0.752) and the semantic task (Spearman’s R = 0.764), closely replicating the results of Study 1. This is shown in the scatterplots in Figures 2 and 3.

Table 4. Summary statistics for Study 2 laterality indices

| Task | Language | Mean trials | mean LI | se LI | % left | % bilateral | % right |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Phonological | L1 | 19.08 | 1.85 | 0.27 | 71 | 29 | 0 |
| Phonological | L2 | 19.5 | 1.92 | 0.38 | 75 | 21 | 4 |
| Semantic | L1 | 18.96 | 1.47 | 0.28 | 71 | 29 | 0 |
| Semantic | L2 | 19.08 | 1.61 | 0.29 | 62 | 33 | 4 |



*Figure 2. Scatterplot showing individual mean LIs in L1 and L2 for Phonological Fluency, with horizontal and vertical denoting standard errors. The solid grey line corresponds to the point of equality of the two measures, and the dotted lines show the limits where difference between LIs is +/- 2.5.*



*Figure 3. Scatterplot showing individual mean LIs in L1 and L2 for Semantic Fluency, with horizontal and vertical denoting standard errors. The solid grey line corresponds to the point of equality of the two measures, and the dotted lines show the limits where difference between LIs is +/- 2.5.*

### Correlations Between LI and Age of Acquisition

We explored whether age of acquisition for English was related to strength of laterality in L2. A significant correlation was found between AOA and LI for the phonological task (r = -0.54, p = 0.039; Figure 2) but not the semantic task (r = -0.34, p = 0.222; Figure 3)

It is worth noting that AOA was only available for 15 of the 25 subjects.

## Discussion

TO BE COMPLETED

# General Discussion